

# From Perception to Cognition: A Survey of Vision-Language Interactive Reasoning in Multimodal Large Language Models

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma<sup>\*</sup>, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, Fei Luo, Tianyi Jiang, Xiaohua Chen, Xiaoshuai Hao, Hehan Li, Andi Zhang, Wenzuan Wang, Kaiyan Zhang, Guoli Jia, Lingling Li, *Senior Member, IEEE*, Zhiwu Lu, *Senior Member, IEEE*, Yang Lu<sup>\*</sup>, *Senior Member, IEEE*, and Yike Guo, *Fellow, IEEE*

**Abstract**—Multimodal Large Language Models (MLLMs) strive to achieve a profound, human-like understanding of and interaction with the physical world, but often exhibit a shallow and incoherent integration when acquiring information (Perception) and conducting reasoning (Cognition). This disconnect leads to a spectrum of reasoning failures, with hallucination being the most prominent. Collectively, these issues expose a fundamental challenge: the ability to process pixels does not yet confer the ability to construct a coherent, credible internal world model. To systematically dissect and address this challenge, this survey introduces a novel and unified analytical framework: “From Perception to Cognition.” We deconstruct the complex process of vision-language interactive understanding into two interdependent layers: Perception, the foundational ability to accurately extract visual information and achieve fine-grained alignment with textual instructions; and Cognition, the higher-order capability for proactive, multi-step, goal-oriented reasoning built upon this perceptual foundation, the core of which is the formation of a dynamic observe-think-verify reasoning loop. Guided by this framework, this paper systematically analyzes the key bottlenecks of current MLLMs at both layers. It surveys the landscape of cutting-edge methods designed to address these challenges, spanning from techniques that enhance low-level visual representations to those that improve high-level reasoning paradigms. Furthermore, we review critical benchmarks and delineate future research directions. This survey aims to provide the research community with a clear, structured perspective for understanding the intrinsic limitations of current MLLMs and to illuminate the path toward building next-generation models capable of deep reasoning and a genuine understanding of the world.

**Index Terms**—Multimodal Large Language Models (MLLMs), Interactive Vision-Language Reasoning, Perception and Cognition

## 1 INTRODUCTION

THE rapid advancement of Multimodal Large Language Models (MLLMs) is propelling the field of artificial intelligence toward its long-standing goal of Artificial General Intelligence (AGI) [1], [2], [3], [4]: creating agents that can perceive, reason, and interact with the physical world in a human-like manner [5], [6], [7], [8]. Central to this progress is the profound synthesis of the sophisticated symbolic reasoning capabilities of Large Language Models (LLMs) [9], [10] with the potent perceptual acuity of Computer Vision (CV) foundation models [11], [12]. On one hand, LLMs such as the GPT series [13], have acquired extensive world knowledge and formidable logical reasoning skills through pre-training on vast text corpora. However, they are inherently confined to a purely symbolic space, operating as “blind” reasoners, detached from the sensory richness of the physical world. Conversely, vision foundation models like CLIP [11] have successfully mapped visual and linguistic

modalities into a unified embedding space [14], [15], enabling unprecedented perceptual generalization. Yet, they typically lack the deep cognitive faculties required for complex, multi-step reasoning.

The emergence of Multimodal Large Language Models (MLLMs) marked the initial exploration into integrating these two capabilities. In this early Exploration Phase (How to Connect?), exemplified by pioneering works such as Flamingo [16] and LLaVA [17], the central challenge was a technical one: how to effectively connect a vision encoder to an LLM. At the time, the research community was primarily focused on solving foundational engineering problems like architectural design and feature alignment, with the primary objective being to make the connection viable. Once the connection was successfully established, and despite progress on numerous benchmarks, the brittleness of these models became exposed when confronted with scenarios demanding fine-grained perception and complex reasoning. This is specifically manifested in the prevalence of issues such as hallucination and bias in state-of-the-art models [18], [19], [20], [21], [22], including Qwen2.5-VL [23], InternVL 2.5 [24], and GPT-4o [25]. These models frequently misinterpret visual details (a perceptual deficit) and are unable to maintain coherent logical chains (a cognitive deficit) [15], [26]. Consequently, the research focus has shifted toward a systematic developmental path from Perception to

---

- Yanbiao Ma is with the Gaoling School of Artificial Intelligence, Renmin University of China.  
E-mail: ybma1998@ruc.edu.cn
- Yang Lu is with Xiamen University, China.
- Yike Guo is with The Hong Kong University of Science and Technology, Hong Kong SAR, China.
- Chenyue Zhou is with Nanyang Technological University, Singapore.

Manuscript received September 28, 2025.

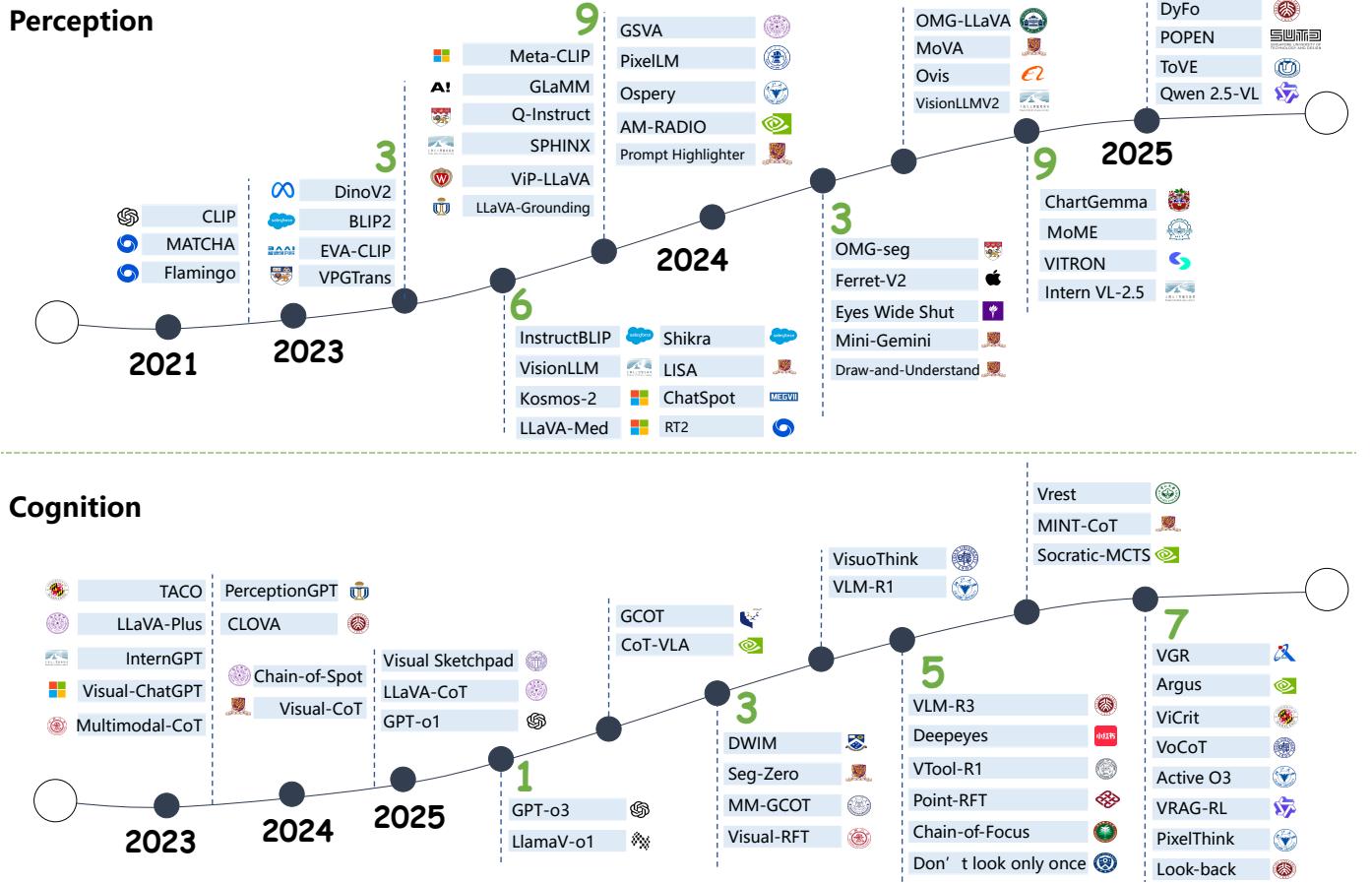


Fig. 1. Evolution of representative multimodal large language models from 2021 to 2025 organized along Perception and Cognition.

**Cognition.** In particular, the approach of first establishing a precise **Perception** layer as a prerequisite for building advanced **Cognition** has gradually become a consensus for enhancing the capabilities of MLLMs [15], [27].

To systematically chart this evolutionary trajectory, this survey introduces the “**From Perception to Cognition**” framework as a unified lens for analysis. It deconstructs the complex process of vision-language interactive reasoning into two distinct yet interconnected layers. The first, Perception, represents the foundational ability to accurately extract visual information and establish fine-grained alignment with textual instructions. This requires not only recognizing objects, attributes, and relations but also precisely grounding textual concepts to specific visual details. The second, Cognition, is the higher-order capability for proactive, goal-oriented, multi-step reasoning built upon this perceptual foundation. It involves decomposing complex problems, planning logical steps, and, critically, the ability to dynamically re-examine visual evidence to validate or refine its reasoning path, thus forming an “**observe-think-verify**” feedback loop. As illustrated in Fig. 1, we present the timeline of MLLM development from 2021 to 2025, organized along these two central pillars of Perception and Cognition.

Leveraging this framework, this survey provides a structured and comprehensive review of the key problems, methodologies, benchmarks, and future directions in interactive vision-language reasoning. Our review begins by analyzing the core challenges at the perceptual and cognitive levels that have driven the field’s progress. Subsequently,

we survey the cutting-edge methods aimed at **enhancing Perception** (e.g., advanced visual encoders) and those designed to **bolster Cognition** (e.g., sophisticated Chain-of-Thought paradigms and dynamic reasoning mechanisms). By adopting this “perception-to-cognition” perspective, we aim to elucidate the limitations of existing models and chart a developmental path for technological evolution.

### 1.1 Pipeline of the Survey

This survey presents a comprehensive overview of the key problems, methodologies, benchmarks, and future directions in interactive vision-language reasoning within Multimodal Large Language Models (MLLMs). As illustrated in Fig. 2, our discussion is structured as follows:

Sec. 2 outlines the fundamental issues in vision-language interaction. We first establish our core analytical framework by defining *perception* (Sec. 2.1) and *cognition* (Sec. 2.2). We then use this framework to analyze the primary challenges that MLLMs face along these two dimensions (Sec. 2.3).

Sec. 3 reviews the evolution of relevant methods, organized according to the challenges they address within our perception-cognition framework. To solve **perception-level** problems, we explore techniques focused on enhancing fine-grained visual capabilities (Sec. 3.1) and improving vision-language alignment (Sec. 3.2). To address **cognitive-level** challenges, we delve into methods for enhancing problem decomposition (Sec. 3.3) and mitigating hallucinations by enabling dynamic reasoning that overcomes the limitations of static perceptual memory (Sec. 3.4).

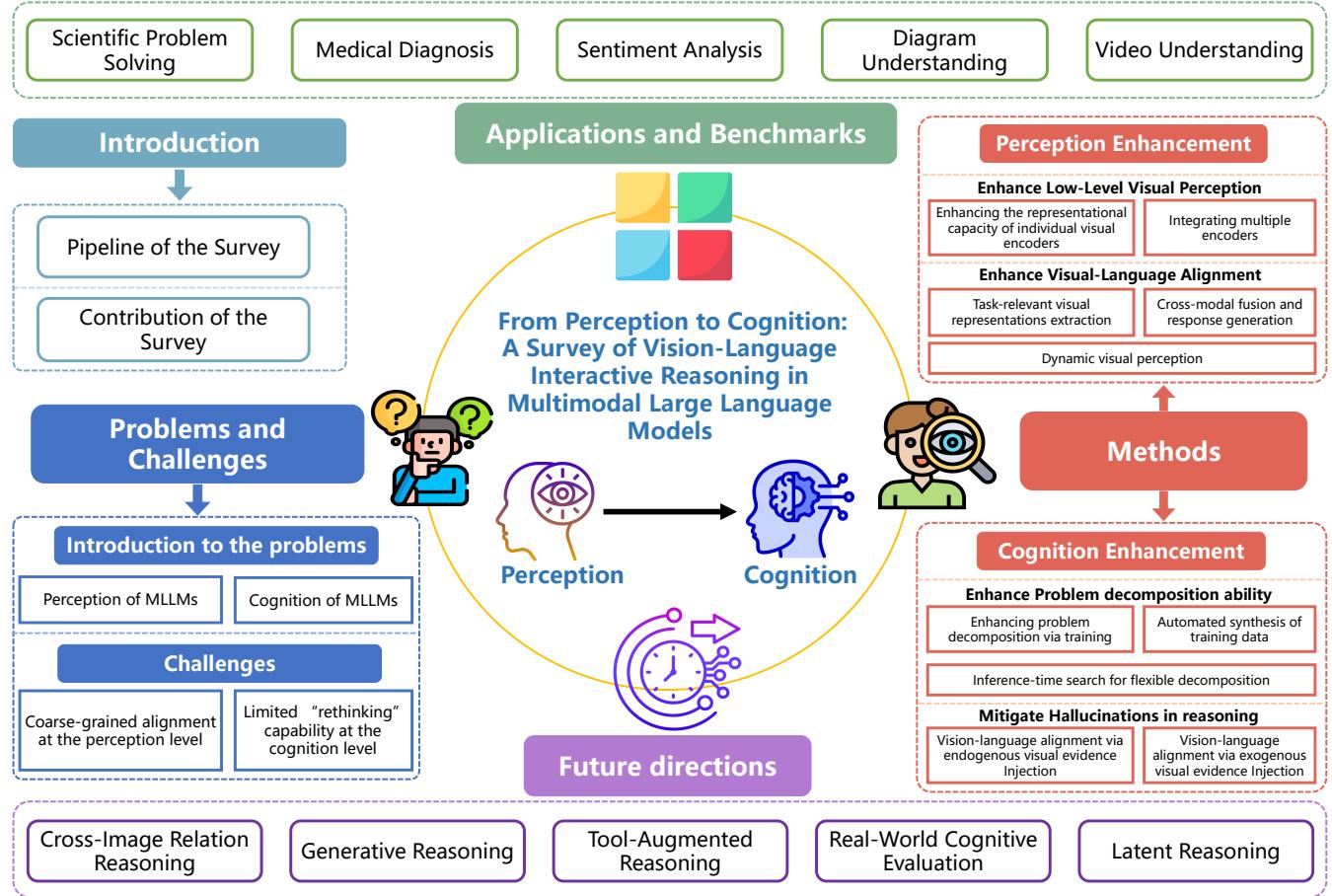


Fig. 2. The overview of the survey structure.

Sec. 4 provides a thorough analysis of key benchmarks and applications across diverse domains, including scientific problem-solving, medical diagnosis, diagram understanding, and video reasoning. This section evaluates how current models perform on tasks that require different balances of perceptual and cognitive skills.

Sec. 5 concludes the survey by discussing promising future research avenues. We explore emerging paradigms such as latent space reasoning, generative reasoning, and tool-augmented reasoning, highlighting how they might further bridge the gap between perception and cognition.

## 1.2 Contribution of the Survey

Recent years have witnessed a surge in survey papers on Multimodal Large Language Models (MLLMs), each providing a unique perspective on this rapidly evolving field. Among the first, Yin *et al.* [28] offered a foundational overview of MLLM development up to early 2024. Subsequent work began to specialize. Some surveys delved into the crucial area of reasoning, analyzing methods for enhancing step-by-step “slow thinking” processes [29], [30]. More recently, a number of surveys [31], [32], [33] has converged on the theme of “thinking with images,” systematically analyzing advancements in fine-grained visual reasoning.

While these surveys provide invaluable insights, they tend to focus either on general MLLM architectures or on specific facets of high-level reasoning. Our survey is distinct

from prior work by adopting a more foundational perspective that deconstructs interactive reasoning into two core, interconnected components: perception and cognition. Based on this approach, we make the following key contributions:

- A Novel Analytical Framework:** We introduce a “Perception-Cognition” framework that provides a structured lens to understand the fundamental challenges in vision-language interaction. This framework moves beyond a surface-level categorization of tasks or models to dissect the root causes of model failures, such as hallucination, by mapping them to specific deficiencies at either the perceptual or cognitive level.
- A Structured Taxonomy of Methodologies:** Based on this framework, we provide a systematic and coherent taxonomy of existing methods. We demonstrate how seemingly disparate research efforts (such as enhancing visual encoders and developing advanced Chain-of-Thought) are, in fact, targeted efforts to solve distinct problems along the perception-to-cognition continuum. This clarifies the relationships between different lines of research.
- A Unified Perspective on the MLLM Developmental Path:** By structuring our analysis along the Perception-to-Cognition axis, this survey explicitly highlights the fundamental dependency of high-level reasoning on the quality of low-level visual representation. It reframes these two domains not as isolated research areas, but as integral and sequential stages in the development of MLLMs. This unified view offers a holistic roadmap, illustrating that

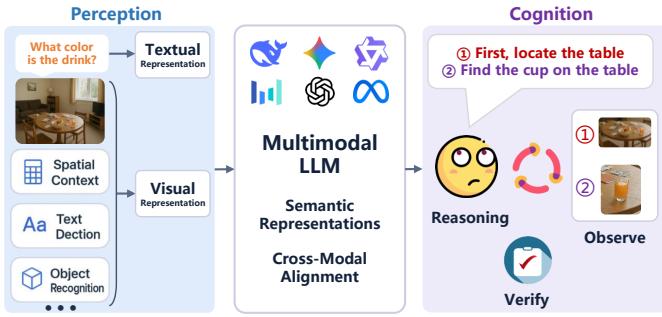


Fig. 3. Overview of the perception–cognition loop. Perceptual modules extract semantic and spatial evidence which are aligned by a multimodal LLM. Cognition then executes a plan–observe–reason cycle with iterative verification to ground each step in visual evidence.

advancements in Perception are a necessary foundation for achieving breakthroughs in Cognition.

## 2 PROBLEMS AND CHALLENGES

Before we deeply explore the classification of methods for vision-language interactive Reasoning in Multimodal large language models (MLLMs), we need to define a core concept to clarify the scope of a model’s capabilities in vision-language interactive understanding: *What are the perception and cognition of MLLMs?* As depicted in Fig. 3, we summarize the relationship between perception and cognition.

### 2.1 Perception of MLLMs

Perception primarily refers to the ability of multimodal large models to accurately extract relevant visual information from input images when tackling specific visual tasks, and to further encode this raw visual data into semantically meaningful visual representations [34], [35], [11]. This process involves not only recognizing visual features such as objects, backgrounds, and texts [36], [37], [38] within images, but also discerning spatial relationships between objects, contextual associations [39], [40] and deep mining of latent semantic information [41]. More critically, effective perception demands that models align these encoded visual representations with input textual information, such as question descriptions, instruction prompts, or dialogue content. This ensures semantic correspondence and mutual reinforcement between visual and textual data. Through such alignment mechanisms, models can provide clear, reliable visual evidence during subsequent vision-language interactions—validating reasoning accuracy or serving as direct answer substantiation. In short, perception constitutes the foundational capability for models to “see” and “understand” visual information, laying the groundwork for subsequent higher-level cognitive operations.

### 2.2 Cognition of MLLMs

Compared with perception, which emphasizes information extraction and alignment, cognition focuses more on a model’s proactive decision-making and reasoning capabilities driven by visual tasks. Specifically, cognition manifests as a model’s ability to systematically determine *when to examine visual information* and *which specific image regions*

*to focus on* based on current task requirements. This selective attention is not random behavior but rather decisions made by the model by integrating existing reasoning processes with visual textual evidence. After acquiring relevant visual information, the model integrates existing textual and visual data to perform further reasoning. Crucially, cognition also requires the model to dynamically assess the sufficiency of current information at each reasoning step: if existing evidence proves insufficient to support conclusions, the model proactively supplements additional visual evidence (e.g., refocusing on relevant image regions to capture finer features); conversely, if current information adequately supports the inference, it directly generates the final answer. This process forms a closed-loop cycle of “observation” (acquiring visual information), “thinking” (multimodal reasoning), and “verification” (validating the match between reasoning results and evidence). This ensures the model’s cognitive process not only maintains logical consistency but also dynamically adjusts its review strategy based on task difficulty, enabling handling of complex vision-language interaction tasks.

### 2.3 Challenges

Building on a clear understanding and distinction of perception and cognition in MLLMs, this section will, within this framework, elaborate on the challenges that MLLMs face in vision-language reasoning. Despite significant advancements in MLLMs for vision-language understanding, where they achieve near-human or even superior performance in tasks like image captioning and visual question answering—a series of critical issues remain unresolved in practical applications. These challenges impose heightened demands on both perception and cognition. **We group these challenges into two primary categories, corresponding to the domains of perception (Sec. 2.3.1) and cognition (Sec. 2.3.2).**

#### 2.3.1 Barriers to Accurate Perception: Weak Extraction and Coarse Alignment

- **Weak Low-Level Visual Information Extraction Capability:** The fundamental limitation of early mainstream MLLM models, such as the LLaVA [17] [42] series, stems from the shortcomings of their underlying CLIP-ViT visual encoders. CLIP-ViT’s pretraining objective focuses on global vision-language alignment, learning matches between the visual semantics of entire images and entire text rather than precise regional or pixel-level alignment. This alignment approach leads to poor fine-grained information recognition and weak spatial localization capabilities. Beyond perceptual limitations, CLIP [11] also suffers from specific data generalization issues. The absence of structured, symbolic visual information (e.g., mathematical geometry problems, charts) in its pretraining dataset causes suboptimal performance in scenarios requiring symbolic or structured understanding.

- **Limited Interaction between Visual and Textual Information:** Most existing MLLMs first encode images independently before projecting them uniformly into the decoder. Cross-modal interaction primarily relies on global relevance alignment, failing to map language queries to specific regions or pixels. Consequently, the coupled information between text and vision is underutilized.

### 2.3.2 Limited Rethinking Capability Constrains Reasoning

- **Challenges in Decomposing Complex Vision-Language Interaction Tasks:** Lacking executable task decomposition paradigms, existing methods predominantly employ single-step or fixed-template approaches, supervising only outcome correctness during training rather than process correctness and consistency.
- **One-Shot Evidence-Based Memory-Based Reasoning Leads to Forgetting and Hallucinations:** A fundamental limitation of most MLLMs is their static, single-pass approach to visual processing during inference. They perform a single encoding of the image at the input stage and do not revisit it during the subsequent reasoning and text generation process. This lack of a dynamic perception-reasoning loop, where visual evidence can be re-examined as needed, often causes a decoupling between the final answer and the underlying visual facts. This issue is particularly exacerbated in complex, long-chain tasks that require understanding the spatial and semantic relationships among multiple visual elements. Consequently, enabling more robust reasoning requires a dynamic mechanism capable of strategically deciding **when** to re-examine the image for evidence, **where** to focus attention, and **what** specific information to extract at each step of the reasoning process.

In summary, by systematically deconstructing the complex challenges of MLLMs into bottlenecks at the **perceptual level** and reasoning limitations at the **cognitive level**, we not only gain a clearer diagnosis of the root causes behind model failures like hallucination but also reveal the underlying logic of the field’s technological evolution. It is precisely these specific challenges that constitute the primary driving force behind advancements in MLLMs, spurring a vast body of work aimed at either bolstering perception or deepening cognition. In the following section, guided by this framework, we will systematically review and analyze the cutting-edge methodologies that the research community has proposed to overcome these distinct perceptual and cognitive challenges.

## 3 METHODS: A SURVEY

In the previous chapter, we systematically analyzed the challenges in MLLMs’ vision-language interaction understanding based on the dimensions of perception and cognition. This chapter will dive into research approaches based on the issues raised in the previous Sec. 3.1 and Sec. 3.2 focusing on addressing perception-related problems, corresponding to the subproblems proposed in Sec. 2.3.1. Sec. 3.3 and Sec. 3.4 concentrate on solving cognition-related problems, corresponding to the subproblems proposed in Sec. 2.3.2.

### 3.1 Enhance the Low-Level Visual Perception of MLLMs

To address the common limitation of weak capacity in low-level visual information extraction, a large number of recent works have concentrated on advancing visual encoders themselves, with the goal of fundamentally improving the representation of fine-grained visual details. We summarize the related methods in Table 1. This line of progress has evolved along two main directions: **enhancing the representational capacity of individual visual encoders** and **integrating multiple encoders in a complementary manner**.

- **Enhance the Representational Capacity of Individual Vision Encoders.** To address the insufficient representation of detail in early encoders like CLIP [11], subsequent research on visual foundation models has focused on two main areas: improving **fine-grained representation** and **geometric-texture representation**. On one hand, to enhance fine-grained representation, subsequent visual encoders such as MetaCLIP [45], SigLip [44], and Eva-CLIP [43] improved semantic alignment and fine-grained recognition abilities by optimizing training objectives and constructing high-quality datasets. On the other hand, to improve geometric-texture representation, several works like DINO series [65], [46], [12] proposed a self-supervised training approach to explore the intrinsic properties of visual data. This enabled them to generate visual representations rich in pixel-level geometric information and texture details, leading to strong performance on tasks requiring sophisticated structural awareness, such as segmentation, localization, and depth estimation. Furthermore, DIVA [47] and VLV [48] unify image generation and understanding by distilling the superior fine-grained representations from generative models into the CLIP-based visual encoder, thereby bridging the gap between generative and discriminative visual understanding.

- **Multi-Encoder Integration and Distillation.** Given that foundational encoders like CLIP [11] excel at high-level semantic representation but lack fine-grained geometric detail, and newer models like DINOv2 [46] provide rich structural and pixel-level representations, research naturally progressed towards combining their complementary strengths. Early explorations like Eyes Wide Out [27], Prismatic VLMs [49], FerretV2 [50], SPHINX [53], MouSi [51], BRAVE [52] proposed a static fusion of features from both CLIP and DINOv2. LLaVA-HR [66] introduces a hybrid-resolution adapter to inject high-resolution features into a low-resolution visual encoder. Mini-Gemini [67] employs CLIP-generated tokens as low-resolution queries, which cross-attend to features from a separate high-resolution encoder within localized windows at corresponding spatial positions. ParGo [54] maps the visual features into Partial tokens, which interact only with a subset of visual features, and Global tokens, which interact with all visual features, thereby enabling effective capture of both local and global information in the image. Recent studies [55] have shown through extensive experiments that directly fusing multi-layer visual features at the input stage achieves more stable and effective performance.

Such fusion mechanisms improve model performance on fine-grained recognition, small object perception, and precise spatial localization—tasks that demand rich and multi-scale visual understanding. However, these approaches rely on static feature integration, which simply stitches together visual representations from different “expert” encoders without adaptive modulation. This rigid combination may lead to conflicting feature requirements across diverse tasks, limiting the model’s flexibility and contextual adaptability in complex vision-language scenarios.

To address this, more advanced methods such as MoME [56], MoVA [57], VisionWeaver [58], TOVE [59], R2-T2 [60] introduced Mixture-of-Experts (MoE) [68] architectures. These models dynamically weight and combine fine-grained and geometric features from different experts based on task requirements, mitigating feature conflicts.

TABLE 1  
Representative vision backbone models and their enhanced low-level visual representations.

Research Direction	Method	Venue	Main Contribution
Single-encoder enhancement	<i>a. Single-encoder optimization.</i>		
	EVA-CLIP [43]	arXiv'23	proposing training methods to enhance the efficiency and stability of CLIP in large-scale settings
	SigLip [44]	CVPR'23	Leveraging a sigmoid loss for enhancing fine-grained representation.
	MetaCLIP [45]	ICLR'24	Enhancing fine-grained representation by training in larger-scale, high-quality datasets.
	DINOv2 [46]	TMLR'25	Strengthening geometric and structural representation through self-supervised learning.
Multi-encoder Integration	DIVA [47]	ICLR'25	By conditioning the diffusion model on dense visual features from CLIP and applying reconstruction loss to optimize CLIP.
	VLV [48]	arXiv'25	Improving CLIP through a unified architecture for image understanding and generation.
	<i>b. Static fusion of encoders.</i>		
	Eyes Wide Shut [27]	CVPR'24	Directly concatenating the visual tokens from CLIP and DINOv2 in an alternating, interleaved manner to form a longer sequence.
	Prismatic VLMs [49]	ICML'24	Concatenating the visual tokens from CLIP and DINOv2 along the channel dimension without increasing the sequence length.
<i>c. MoE-based multi-encoder fusion.</i>	Ferret-v2 [50]	CoLM'24	Employing multi-granularity fusion of visual representations for fine-grained perception.
	MouSi [51]	CoLM'24	Integrating multi-scale encoder outputs, combining global context with local structural cues for robust spatial reasoning.
	BRAVE [52]	ECCV'24	Statically fusing visual features by adopting resolution-aware projection to preserve both high-resolution local detail and low-resolution global semantics.
	SPHINX [53]	ECCV'24	Leveraging hierarchical fusion of CLIP and DINO to balance semantic alignment and geometric precision.
	ParGo [54]	AAAI'25	Introducing local and global tokens with dedicated attention masks to extract both fine-grained and holistic visual information.
<i>d. Distillation-based models.</i>	Layer_Select_Fuse [55]	CVPR'25	Directly fusing multi-layer visual features at the input stage proves to be more stable and effective.
	MoME [56]	NeurIPS'24	Task-adaptively fusing CLIP and DINOv2 features via a Mixture-of-Experts.
	MoVA [57]	NeurIPS'24	Enhancing cross-task representation flexibility using a multi-expert MoE.
	VisionWeaver [58]	EMNLP'25	Dynamically combining SAM, Vary (specialized in text recognition) and DINOv2 through a routing module.
	TOVE [59]	ICLR'25	Feature combination is achieved through coarse-grained context-aware expert routing and a fine-grained expert fusion module.
	R2-T2 [60]	ICML'25	Locally optimizing the routing weights at test time by steering them toward those of correctly predicted samples in the neighborhood.
	Radio [61]	CVPR'24	Enabling a single encoder to absorb heterogeneous vision encoders capabilities via multi-teacher feature distillation.
	UNIC [62]	ECCV'24	Enabling a universal classifier to inherit cross-task skills from multiple teachers encoders via multi-teacher, layer-wise distillation.
	MoVE-KD [63]	CVPR'25	Enabling a single encoder to inherit the distinct abilities of multiple experts via distillation.
	DUNE [64]	CVPR'25	Compressing multi-expert knowledge into an efficient and powerful lightweight model.

While effective, using multiple encoders drastically increases computational costs. As a solution, works such as MoVE-KD [63], Radio [61], UNIC [62] and Dune [64] have employed knowledge distillation. This approach transfers the combined strengths of multiple expert “teachers” into a single, efficient “student” encoder. The resulting student model can generate a unified representation that effectively captures both fine-grained details and geometric structure, achieving performance comparable to multi-encoder models but at a significantly lower computational cost.

### 3.2 Enhance Vision-Language Alignment in MLLMs

While prior work has significantly improved the general-purpose representation capabilities of visual encoders, a new core focus has emerged in the field of MLLMs: **enhancing**

**task-driven vision-language alignment for better interactive understanding.** The core concept of this alignment is to effectively connect task-agnostic, general visual semantics, encompassing broad information about objects, attributes, and relations, with the specific goals and constraints contained in a user’s instruction. This instruction-guided workflow transforms general-purpose visual representations into task-relevant ones through two primary stages:

- **Task-Relevant Visual Representations Extraction.** Conditioned on the instruction’s semantics, the model selectively extracts relevant representations from the high-dimensional, general-purpose visual representation space (Sec. 3.2.1).
- **Cross-Modal Fusion and Response Generation.** The extracted visual representations are then deeply fused with the instruction’s semantic representation to generate a pre-

TABLE 2  
Representative methods grouped by research directions. We keep ultra-brief contributions for compactness.

Research Direction	Method	Venue	Main Contribution
Task-Relevant Visual Representations Extraction	(a) Improve the projection layer.		
	Honeybee [69]	CVPR'24	Proposing a flexible and locality-preserving visual projector that balances efficiency with spatial understanding.
	Uni-Med [70]	NeurIPS'24	Proposing a CMoE module to manage inter-task relationships via dynamic resource allocation.
	ChartMoE [71]	ICLR'25	Proposing a MOE projector for aligning textual and graphical elements in different charts.
	LLaVA-ST [72]	CVPR'25	Proposing a post-processing module that compresses projected video embeddings while preserving their spatio-temporal relationships.
	LLaVA-Octopus [73]	arXiv'25	Proposing an instruction-driven adaptive Projector.
	Ovis2.5 [74]	arXiv'25	Proposing a visual embedding table to replace the MLP projection network.
	(b) Task-specific fine-tuning.		
	MATCHA [75]	ACL'23	Math-centric SFT
	LLaVA-Med [76]	NeurIPS'23	Medical SFT
	Q-Instruct [77]	CVPR'24	Low-level perception SFT
	ChartInstruct [78]	ACL'24	Chart SFT
	(c) Prompt-tuning.		
	VPT [79]	ECCV'22	Proposing visual prompts for adapting downstream visual tasks
	VPGTrans [80]	NeurIPS'23	Proposing transferable visual prompts generator across MLLMs
	TVP [81]	CVPR'24	Proposing joint learnable visual and text prompts for coordinated task adaptation
Cross-Modal Fusion and Response Generation	(d) Improve instruction encoding paradigm.		
	Shikra [82]	arXiv'23	Encoding coordinates in text instruction.
	LLaVA-Grounding [83]	ECCV'24	Encoding boxes and masks in prompts, unifying visual chat with detection
	Kosmos-2 [84]	ICLR'24	Integrating ground tokens into vocabulary, allowing the model to generate bounding boxes as its instruction-driven output.
	GLaMM [85]	CVPR'24	Introducing mask token to vocabulary, enabling the model to predict the pixel-level mask
	ViP-LLaVA [86]	CVPR'24	Encoding visual instruction prompts via visual feature fusion
	Draw-and-Understand [87]	ICLR'25	Proposing a visual instruction prompt encoder
	(e) Enhance output architecture.		
	LISA [88]	CVPR'24	Introducing <SEG> token to generate decoder-based segmentation outputs
	GSVA [89]	CVPR'24	Introducing <SEG> token for multi-targets mask and <REJ> token to reject empty targets
	VisionLLM v2 [90]	NeurIPS'24	Introducing unified multi-task decoders
	VITRON [91]	NeurIPS'24	Proposing an unified paradigm for video and image segmentation.
Dynamic Perception	M2SA [92]	ICLR'25	Proposing early visual feature fusion and multiple <SEG> tokens to generate more fine-grained output
	POOPEN [93]	CVPR'25	Proposing preference-based segmentation
	(d) Visual search.		
	V* [94]	CVPR'24	Iterative look-back search
	DyFo [95]	CVPR'25	MCTS-guided focus
	FaST [96]	ICLR'25	Fast or slow visual search

cise, structured output that meets the task’s requirements (Sec. 3.2.2).

However, the methods in these stages represent a static, single-turn form of interaction. To address this limitation, a key research direction has been to build a *dynamic perception* mechanism (Sec. 3.2.3). The goal is to give the model the ability for active perception. Specifically, the model can evaluate its current understanding to determine if more visual evidence is needed, and then actively re-examine the image to get that evidence. This creates an iterative visual search loop.

### 3.2.1 Task-Relevant Visual Representation Extraction

Following the aforementioned framework, we first explore the core strategies for the first stage: enhancing the model’s ability to extract task-relevant visual representations. Researchers have primarily proposed three main approaches: **improving the projection layer**, **task-specific fine-tuning**, and **prompt tuning**. These methods focus on strengthening the expression of task-specific visual features by optimizing

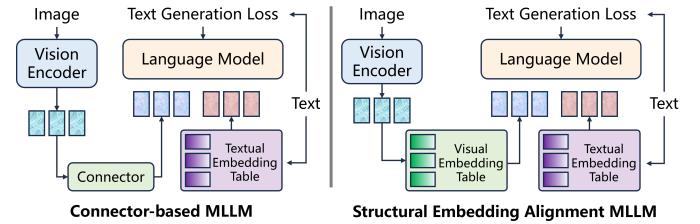


Fig. 4. Left: Connector-based MLLM: The typical architecture of traditional multimodal models (e.g., LLaVA), where the connector is usually an MLP that projects visual features into the same dimensional space as text embeddings. Right: Structured embedding alignment in Ovis: The output of the visual encoder is no longer directly projected through an MLP, but instead mapped to a visual embedding table.

the alignment and adaptation mechanism that maps visual features into the language space, all without altering the underlying visual backbone. As illustrated in Table 2, we categorize the representative works in task-relevant visual representation extraction.

- **Improve the Projection Layer.** As the bridge connecting

the visual encoder and the large language model, the design quality of the projection layer directly determines the depth and precision of the model’s visual understanding. Traditional static projection methods, such as a simple MLP [97], indiscriminately map all visual features, creating a significant information bottleneck [16], [98], [99]. Therefore, the core significance of improving the projection layer lies not only in transmitting richer visual information but, more importantly, in empowering it with the ability to dynamically extract and transform relevant visual features based on the specific task.

For instance, in the fields of chart understanding and medical question answering, ChartMoE [71] and Uni-Med [70] employ a MOE [68] architecture. This structure consists of multiple identical two-layer MLPs and uses a Top-k routing mechanism to weight and combine their outputs based on the input content. In video understanding, LLaVA-Octopus [73] adaptively fuses the outputs of multiple specialized projectors according to the given instruction. To enhance the model’s fine-grained perception of spatial relationships, LLaVA-ST [72] and Honeybee [69] respectively reinforce the spatial properties of visual features before projection by explicitly modeling local relationships and introducing space-aware convolutions. This is crucial for localization and referring tasks. ParGo [54] proposes an innovative global-local projector that bridges vision and language by integrating both global context and local details, overcoming the over-focus on salient regions in traditional methods, enabling more comprehensive and fine-grained visual representation, while effectively controlling computational cost through token length management, thus achieving efficient alignment between visual features and the LLM. As shown in Fig.4, the Ovis series [74], [100] no longer projects the output of the visual encoder through an MLP, but instead maps it into a learnable visual embedding table for transformation. This structure is similar to the text embedding table.

- **Task-Specific Fine-Tuning.** Instruction fine-tuning on task-specific vision-language datasets, such as LLaVA-Instruct-150K [17] and MathV360K [101], can enhance a model’s ability to extract visual information tailored to those tasks. For example, MATCHA [75] focuses on mathematical reasoning; Q-Instruct [77] is dedicated to strengthening foundational visual perception for question answering; ChartGemma [102] and ChartInstruct [78] concentrate on the comprehension of chart-based problems; and LLaVA-Med [76] targets visual question answering in the medical domain.

- **Prompt Tuning.** Although instruction fine-tuning is effective, it has limited scalability and typically requires adjusting a large number of parameters, resulting in high training costs. To address this issue, Parameter-Efficient Fine-Tuning (PEFT) techniques have been introduced to the multimodal domain. The core idea originates from Visual Prompt Tuning (VPT) [79] in the computer vision field. This method adds a small number of learnable prompt parameters to the input of a frozen backbone network, enabling adaptation to downstream tasks at a very low cost. Inspired by this, works such as VPGTrans [80] and TVP [81] have designed visual prompts. Without compromising the model’s general-purpose visual representations, these prompts enhance the alignment between task-specific semantics and visual features, thereby achieving efficient and transferable task adaptation.

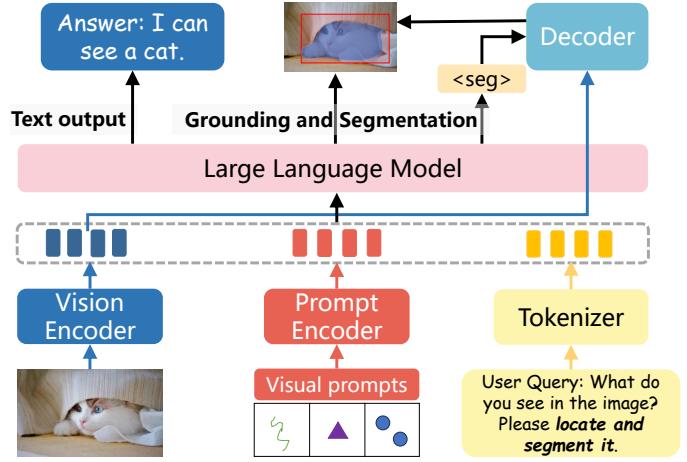


Fig. 5. An illustration of cross modal fusion and response generation. In this figure, the prompt encoder improves the instruction encoding paradigm. The segmentation decoder enables the model output the segmentation mask, which enhances the response generation.

### 3.2.2 Cross-Modal Fusion and Response Generation

Once task-relevant visual representations are obtained, the subsequent challenge is to effectively align these representations with the semantics of the instruction to drive the generation of precise responses. In mainstream MLLM architectures, this cross-modal alignment process primarily relies on cross-attention mechanisms. Consequently, most current research is evolving along two parallel technical paths: The first path focuses on **improving the instruction encoding paradigm**, enabling the model to respond more precisely to specific semantic concepts within the instruction during the alignment process [83]. The second path, by **enhancing the output architecture**, aims to materialize the cross-modal understanding formed during fusion into responses that go beyond the scope of text [103]. These two paths are complementary, sharing the common goal of achieving more fine-grained vision-language alignment. As depicted in Fig. 5, we demonstrate the process of how to enhance the encode paradigm and response generation.

- **Improve Instruction Encoding Paradigm.** To enhance the region-level referring capabilities of MLLMs, the instruction encoding paradigm has undergone a series of developments. Early models, such as Shikra [82] and ChatSpot [115], directly encoded the continuous coordinates of a Region of Interest (ROI) as part of the text sequence. To facilitate model learning, Kosmos-2 [84] first encodes the image after dividing it into  $p \times p$  patches, and then discretizes and normalizes the continuous ROI coordinates, mapping them to positional tokens within the vocabulary. Subsequent research, including GLaMM [85] and Osprey [116], adopted binary masks as input, which provide greater spatial detail, to achieve more precise regional guidance. More recent research has shifted towards more flexible forms of visual instructions: ViP-LLaVA [86] extended the input to include hand-drawn visual markers [117], while Draw-and-Understand [87] and OMG-Seg [118] introduced an external visual prompt encoder. This allows the model to decode the user’s intent from the visual channel without compromising its original global understanding capabilities.

- **Enhance the Output Architecture.** To reduce the ambiguity of text-only outputs in complex visual scenes, researchers

TABLE 3  
An overview of CoT-focused training paradigms and the core issues they aim to address.

Method	Venue	Training	Core Problem
Let’s Verify Step-by-Step [104]	ICLR’24	SFT	Outcome-based supervision cannot guarantee the logical correctness of the problem decomposition process itself.
Multimodal-CoT [105]	ICLR’24	SFT	Text-only reasoning decomposition decouples from visual facts.
ChartGemma [102]	ACL’24	SFT	General MLLMs underperform on charts problems decomposition.
Dolphins [106]	ECCV’24	SFT	Unlocking transferable reasoning capabilities for autonomous driving through fine-tuning.
Visual CoT [107]	NeurIPS’24	SFT	The individual steps of a problem decomposition are not explicitly grounded in verifiable local visual evidence.
LLaVA-CoT [108]	arXiv’24	SFT	Problem decomposition is an emergent ability from prompting, not an innate and robust skill of the model.
LlamaV-o1 [109]	ACL’25	SFT	Whether CoT enhances the adversarial robustness of MLLMs and how reasoning process behaves under adversarial attacks.
Learning Theorem Rationale [110]	AAAI’25	SFT	When decomposing math problems, the reasoning processes of MLLMs fail to adhere relevant mathematical theorems.
UV-DPO [111]	arXiv’25	DPO	Learning to ground each step of a visual problem decomposition via SFT is data-hungry and costly.
Visual Grounded Reasoning [112]	arXiv’25	SFT	The model’s decomposition path is biased by language priors instead of being driven by fine-grained visual evidence.
VLM-R1 [113]	arXiv’25	SFT+GRPO	SFT is insufficient for finding optimal decomposition policies
Visual-RFT [114]	arXiv’25	SFT+GRPO	It is difficult to define effective reward signals to guide the decomposition of complex visual problems via Reinforcement Learning.

have explored output architectures capable of generating pixel-level localization information. Models like LISA [88], LLM-Seg [119], and OMG-LLaVA [120] introduced a special token <SEG>, into the vocabulary. The hidden state vector corresponding to this token is then used as a query to drive a separate segmentation decoder to generate a mask. This paradigm was further extended by works like GSVA [89], MMR [92], Instruction-guided-masking [121] to handle multi-object and multi-granularity segmentation tasks. To address the issue of semantic bias in segmentation results, POPEN [93] employed preference learning to fine-tune the model and suppress incorrect segmentations. Concurrently, to reduce the reliance on large-scale, pixel-level annotated data, works like Llafs [122] and Prompt Highlighter [123] have investigated instruction-based segmentation under few-shot or zero-shot conditions. Furthermore, several works [90], [91], [124], [125] have focused on building a unified output interface. This enables a single model to generate different forms of localization results, such as bounding boxes or segmentation masks, according to the user’s instruction, thereby accommodating diverse task requirements.

### 3.2.3 Dynamic Perception

Building upon the static, single-turn vision-language interaction paradigm discussed previously, this section focuses on the implementation of dynamic perception. The core of these methods is to endow the model with the ability to actively and iteratively search for visual information, thereby overcoming the limitations [126] of static perception.

For example, V\* [94] employs an LLM-guided hierarchical visual search. At inference time, it collaborates with an MLLM to progressively zoom in and look back for evidence, achieving more fine-grained visual question answering in high-resolution and crowded scenes. Subsequent research has focused on reducing training costs and improving architectural coupling [95], [96], [127]. DyFo [95] formalizes visual search as a Monte Carlo Tree Search (MCTS) [128], interacting with external visual experts to dynamically focus on key regions in a training-free setting. Taking a step further,

FaST [96], inspired by the concept of fast and slow thinking, trains a lightweight, built-in adapter to control the speed of reasoning based on the problem’s difficulty, enabling human-like dynamic visual search.

### 3.3 Enhance Problem Decomposition Ability of MLLMs

Early MLLMs operated on a single-step reasoning paradigm, the core deficiency of which lies in treating any task as a monolithic “input-output” mapping process, without any explicit problem decomposition. Consequently, these models often falter when faced with complex reasoning tasks [129]. To overcome this limitation, research has focused on endowing models with the ability to perform step-by-step problem decomposition [130], [131]. The goal is not merely to improve the accuracy of the final answer, but also to ensure the correctness and verifiability of the reasoning process.

Presently, most approaches do not employ a built-in, structured framework for problem decomposition; instead, they rely on prompting techniques, such as Chain-of-Thought (CoT), to implicitly steer the model along a decompositional reasoning path. However, a sole reliance on prompting [105], [132] is insufficient for internalizing this decompositional ability within the model’s parameters. The central research challenge is therefore to transform this decompositional reasoning from a transient, prompt-induced behavior into an innate, self-directed capability of the model. To this end, the academic community [133] is primarily exploring the following three research directions:

- **Enhancing problem decomposition via training.** Dedicated training paradigms can enhance the problem decomposition ability of Multimodal Large Language Models (MLLMs), aiming to make their reasoning more visually grounded and logically sound. (Sec. 3.3.1).
- **Automated Synthesis of Training Data.** To overcome the high cost and scalability issues of manual annotation, researchers have developed automated methods for constructing large-scale, high-quality Chain-of-Thought (CoT) datasets with interleaved visual and textual evidence. (Sec. 3.3.2).

- **Inference-Time Search for Flexible Decomposition.** To overcome the limitations of traditional Chain-of-Thought (CoT), which follows a single, linear reasoning path and risks finding suboptimal solutions, researchers are adapting inference-time search algorithms to explore multiple reasoning paths and find the best possible answer. (Sec. 3.3.3).

### 3.3.1 Enhancing Problem Decomposition via Training

The paradigm of enhancing problem decomposition through process supervision originated in the language-only domain [104]. When this paradigm is migrated to multimodal scenarios, researchers must address a series of new challenges. The primary issue is that the model's decomposition steps must be guided and constrained by visual information to ensure factual consistency. Secondly, for specialized domains such as mathematics, the model must acquire specific decomposition structures that align with the field's intrinsic logic. In Table 3, we present some representative methods to give a first-look understanding of prevailing training paradigms. These methods can be categorized into three main training paradigms:**imitation learning, curriculum learning and preference learning.**

- **Imitation Learning.** Early CoT methods established a foundational framework for problem decomposition, but their generated text-only reasoning paths were prone to decoupling from visual facts [132]. As the entire reasoning process unfolded solely at the textual level, the decomposition plan could be flawed from the outset. Consequently, the core of subsequent research has been to introduce strict constraints to the decomposition process, ensuring that each step of the reasoning path and its required evidence remain consistent with the visual facts. Multimodal-CoT pioneered a two-stage generation process to achieve implicit supervision of visual evidence [105]. The model first generates a textual reasoning chain, which then serves as the condition for producing the final answer. This sequential dependency creates an implicit constraint, compelling the initial reasoning to be visually grounded to ensure the final answer's accuracy. Building on this, to enhance the verifiability of decomposition steps, works like Visual CoT and Visual Grounded Reasoning supervised the model to cite visual evidence such as bounding boxes, transforming ambiguous linguistic references into precise, verifiable coordinate localizations [107], [112]. Furthermore, to improve the domain-specific decomposition logic of MLLMs, works such as ChartGemma [102], Dolphins [106], ChartInstruct [78], Sce2DriveX [134] and Learning Theorem Rationale [110] have fine-tuned models on specialized CoT datasets. This ensures the model's reasoning process adheres to the rigorous paradigms of disciplines like mathematics, thereby validating the effectiveness of its decomposition.

- **Curriculum Learning.** More recently, LLaVA-CoT [108] and LlamaV-01 [109] have proposed a "de-prompted" Supervised Fine-Tuning (SFT) paradigm that introduces the concept of curriculum learning to cultivate the model's decomposition and reasoning abilities through a phased, easy-to-hard process. The model's training unfolds in three stages: it first learns to decompose complex problems into sub-problems, then grounds each sub-problem in visual evidence, and finally integrates this evidence to form a coherent,

summary answer. As a result, these models are trained to autonomously decompose problems at inference time and generate intermediate steps and the final answer accordingly. This curriculum-based training paradigm helps the model form a more robust and innate reasoning capability. However, the effectiveness of SFT is limited by its dependence on imitating a single "correct" decomposition path. This method does not equip models to evaluate alternative strategies or self-correct their reasoning, which limits their generalization in tasks that require flexible, exploratory reasoning. However, the effectiveness of SFT is limited by its dependence on imitating a single "correct" decomposition path. This method does not equip models to evaluate alternative strategies or self-correct their reasoning, which limits their generalization in tasks that require flexible, exploratory reasoning.

- **Preference Learning.** The core idea of preference learning is to learn a predictive model from feedback expressed as relative preferences. This model aims to capture and internalize the underlying judgment criteria hidden within the feedback, thereby enabling it to make preference judgments on unseen instances that align with those criteria. Unlike traditional imitation learning, which provides a positive label for each sample, datasets for preference learning are composed of preference relationships in forms such as comparisons, rankings, or ratings. From a policy optimization perspective, mainstream methods can be divided into on-policy and off-policy learning.

For **on-policy learning**, the core principle is that the data used for policy updates must be generated by the current version of the policy being optimized. This paradigm ensures an unbiased update direction and is generally more stable. Mainstream methods adopt Generalized Reward Policy Optimization [135] (GRPO) and its variants. For example, methods like VLM-R1 [113], Visual-RFT [114], Reason-RFT [136], Seg-Zero [137], R1-OneVision [138] and RAGEN [139] transform visual evidence from the reasoning process into verifiable training signals, thereby enhancing the model's ability to select better problem decomposition paths.

For **off-policy learning**, the core principle is that policy updates can utilize historical experience data generated by any previous policy. It significantly improves data efficiency by correcting for distribution shifts using techniques like importance sampling. Mainstream methods adopt DPO [140] (Direct Preference Optimization) and its variants. For instance, UV-CoT [111] generates multiple chains of thought for the same visual problem, constructs preference pairs based on visual alignment, step-by-step consistency, and the correctness of the final answer, and then updates the policy. V-DPO [141] uses a large number of synthetically generated image preference pairs to enhance the model's inclination to follow visual evidence when decomposing problems during reasoning. VTS-DPO [142] trains a general-purpose verifier/scorer offline using preference pairs from multi-step trajectories to stably evaluate and guide multi-step visual reasoning at inference time.

### 3.3.2 Automated Synthesis of Training Data

The effectiveness of supervised learning methods in enhancing a model's problem decomposition ability is highly dependent on large-scale, high-quality CoT datasets. Early

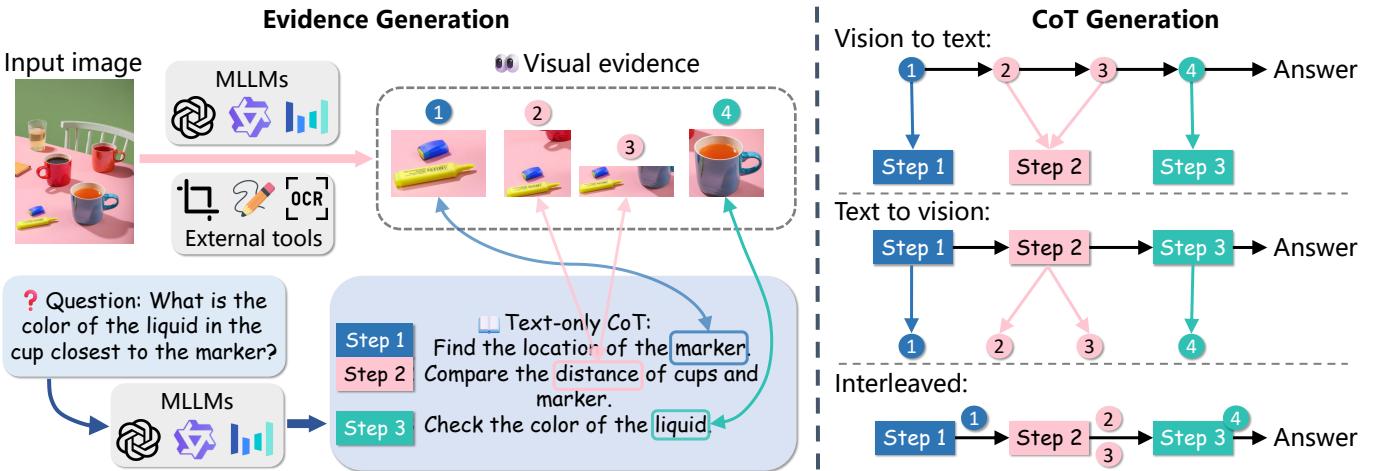


Fig. 6. An example of how to construct an interleaved CoT. We demonstrate three typical ways in the image. The main difference is that, interleaved method generates text rationales and visual evidence in a coherent way, while another two generate text and visual evidence respectively.

benchmarks, such as ScienceQA [143] and A-OKVQA [144], provided detailed, manually annotated reasoning steps and answers for each problem. However, manual annotation is costly, has limited scalability, and struggles to meet the data volume requirements of current training paradigms. Furthermore, this approach often lacks visual evidence grounding for the intermediate steps, making it ineffective at enhancing a model’s ability to decompose specifically visual problems. To mitigate these issues, researchers have adopted automated data synthesis methods. It is important to note that “synthesis” here is defined broadly, encompassing both the generation of entirely new CoT [132] data from scratch for specific tasks, as well as the reprocessing and enhancement of existing datasets which mainly rely on augmenting text-only reasoning processes with visual evidence. Based on the methods for generating and processing CoT data, these approaches can be categorized into two primary strategies: **generation via external teacher models** and **bootstrapped data generation**.

- **Generation via External Teacher Models.** This paradigm primarily employs a teacher model to automate the generation of CoT datasets. Its core task is to decompose a vision-language reasoning problem step-by-step, providing both the correct reasoning process and the final answer. However, the key challenge lies in ensuring that the decomposition path generated by the model is not only logically sound at the textual level but also tightly aligned with the visual evidence. To address this challenge, researchers have explored various synthesis strategies. As illustrated in Fig. 6, these strategies can be categorized into three mainstream approaches based on the sequence in which the textual decomposition steps and the visual evidence are generated and verified: **Vision-to-Text**, **Text-to-Vision** and **Interleaved Vision-Text Generation**.

**(a) Text-to-Vision.** The Text-to-Vision (T2V) paradigm first constructs a text-only reasoning sequence and subsequently matches each textual reasoning step with corresponding visual evidence to build the interleaved reasoning dataset. For example, Cogcom [145] first uses a teacher model to generate a reasoning process that contains placeholders for visual operations. It then executes these operations

to populate the visual results and searches the resulting branching tree with a DFS [146] algorithm to retrieve the optimal reasoning path. The data construction process for Mint-CoT [147] is based on the existing Mulberry-260K [148] dataset. It begins by filtering for mathematical problems and their text-only solutions. Next, it utilizes Optical Character Recognition (OCR) to extract key visual tokens from the corresponding images. Finally, a teacher model is employed to precisely align each textual reasoning step with the visual tokens upon which it depends. However, in this construction paradigm, the generation processes for textual reasoning and visual evidence are entirely decoupled, which can easily introduce hard-to-correct visual hallucinations at the source.

**(b) Vision-to-Text.** This paradigm employs a **vision-first** approach, where a textual rationale is generated only after visual evidence has been established. Representative works like MM-GCoT [149] and SIFTthinker [150] simulate the human viewing pattern by proceeding from a global scan to local details to identify and structure this evidence. Similarly, Pixel-Reasoner [151] first uses a teacher model to locate key visual cues and obtain local views before generating its analysis. In these methods, the process is an evidence-driven “reverse induction”: the textual reasoning is retroactively generated to describe and connect the pre-selected visual cues. While this approach effectively eliminates factual hallucinations by design, its fundamental drawback is that the resulting decomposition is not a top-down problem-solving strategy. Instead, its structure is dictated by the sequence of visual observations, which can lead to a reasoning path that is unnatural, inefficient, or logically suboptimal, despite being factually correct at every step.

**(c) Interleaved Vision-Text Generation.** The joint synthesis of textual reasoning and visual evidence represents a less prevalent research direction. For instance, approaches like LATTE [152] and TACO [153] adopt a collaborative paradigm between a teacher model and multiple external visual tools, in which the teacher model is responsible for decomposing complex problems and planning the reasoning steps. When needed, it calls upon specific visual tools to acquire precise visual evidence. The results returned by these tools are then fed back into the reasoning process, creating a “think-act-

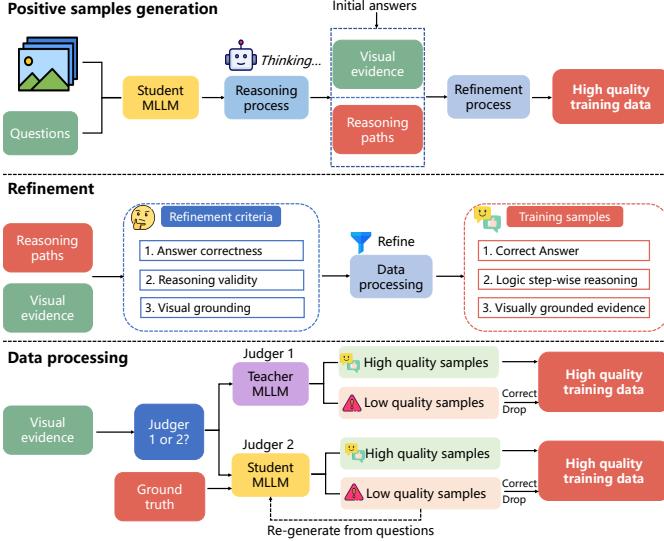


Fig. 7. An illustration of positive samples refinement.

observe” loop that ultimately generates an interleaved vision-language Chain-of-Thought.

- **Bootstrapped Data Generation.** To reduce the reliance on external teacher models, some works employ bootstrapping methods to generate CoT data. The core mechanism is to leverage the model’s own capabilities to generate diverse reasoning paths for a given problem. These paths are then classified as positive or negative samples. This classification is based on the quality of their final outcomes or the logical validity of their reasoning process, with the evaluation performed either by the model itself or by external models. To ensure a rich set of negative examples, some methods also proactively generate them by intentionally injecting errors into the input or the reasoning chain. Through training on such curated data, the model learns to discriminate among candidate reasoning trajectories and prefer those yielding correct, evidence-aligned solutions. Based on whether the generated training data includes negative samples, we classify the methods into two categories: **Positive Sample Refinement** and **Preference Data Generation**. As illustrated in Fig. 7, we demonstrate the generation of positive samples.

**(a) Positive Sample Refinement.** In the text-only domain, early research such as STA-R [154] proposed a method for dynamically augmenting datasets. In this process, the model first generates decomposition paths for a problem based on few-shot examples. If a path proves effective and yields the correct answer, it is considered a high-quality training sample. If not, the model is guided to reverse-engineer a rational decomposition strategy from the correct answer. These corrected and validated high-quality samples are then fed back into the training set, creating a loop of continuous dataset optimization and simultaneous improvement of the model’s capabilities.

In the multimodal domain, this dataset construction strategy has been adapted to generate more complex reasoning samples that contain visual evidence. For instance, MC-CoT [155] constructs training data via bootstrapped filtering to improve the consistency [156] of smaller models in problem decomposition. It first prompts the model to self-iterate and generate multiple candidate decomposition

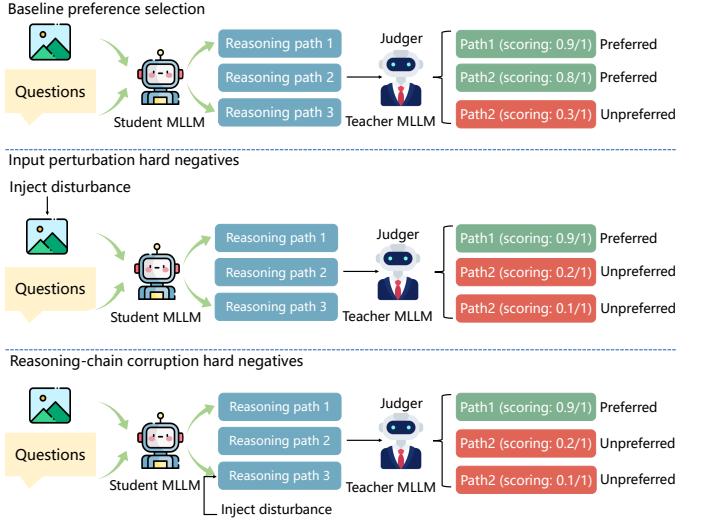


Fig. 8. An example of how to construct preference data. We obtain negative training samples by injecting perturbations into either the inputs or the candidate CoT outputs; alternatively, negatives can be collected without explicit perturbations.

paths. It then uses self-evaluation and a consistency voting mechanism to filter for the solution with the strongest consensus, which is then used as a positive sample for contrastive training against the correct label. GCOT [157] demonstrates how to construct an interleaved vision-language Chain-of-Thought dataset with limited data. This method begins by decomposing a complex problem into a series of initial “subproblem, bounding box” pairs to form a seed dataset. The model then trains on this dataset while continuously filtering and optimizing, retaining only the bounding boxes that best match the visual facts, ultimately producing a complete interleaved CoT dataset.

**(b) Preference Data Generation.** Beyond generating only positive samples, other works focus on the automatic construction of preference pairs for offline learning methods like Direct Preference Optimization (DPO) [140]. As is depicted in Fig. 8, we give an example of how to construct preference data. This is primarily achieved through two strategies: judge-based scoring and bootstrapping. The former uses an external or base model as a “judge” to score multiple candidate responses. This is implemented by either evaluating the final outcomes of different reasoning paths [158], [159], [160] or by using the model’s capacity for self-correction to create positive and negative pairs [161]. The latter strategy, bootstrapping, proactively generates negative samples by injecting perturbations. For example, BPO [162] perturbs the image input or injects logical errors into the reasoning chain to automatically create valuable negative samples.

### 3.3.3 Inference-Time Search for Flexible Decomposition

The traditional CoT paradigm guides a model along a single, linear reasoning path. This process is essentially a greedy search strategy, which risks converging on a locally optimal path while missing the globally optimal solution. To overcome this limitation, methodologies originating from the text-only domain, designed to enhance the flexibility of the reasoning process, are being progressively adapted for multimodal tasks.

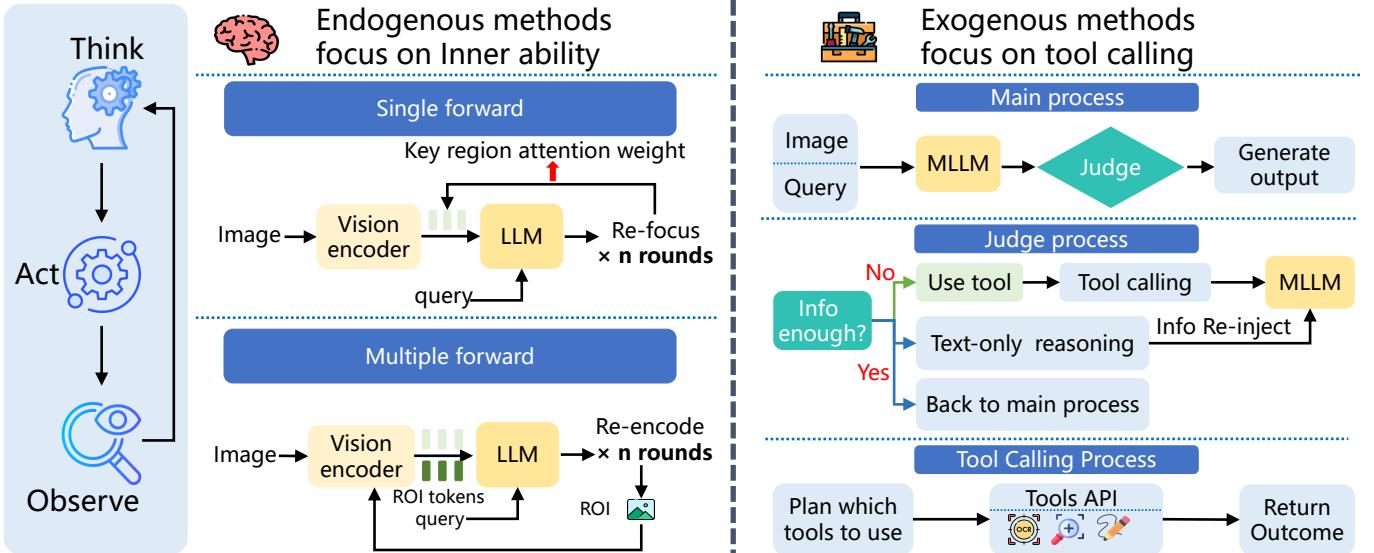


Fig. 9. An illustration of thinking with image.

These methods replace the single, deterministic generation process by exploring a broader solution space. Foundational work such as Tree of Thoughts (ToT) [163] conceptualizes the reasoning process as a tree, where each node represents a partial solution or an intermediate thought. It systematically explores this tree using algorithms like Breadth-First Search (BFS) [164] or Depth-First Search (DFS), enabling backtracking and the exploration of multiple reasoning paths. Building on this foundation, a series of advanced decoding strategies [156], [165], [166], [167] have been proposed, which employ diverse sampling and voting mechanisms to identify the most robust answer. More recently, techniques like AlphaMath [168], Rest-mcts\* [169], and SVPO [170] model the reasoning process as a Monte Carlo Tree Search (MCTS) problem, allowing the model to dynamically allocate more computational resources to the most promising branches of the reasoning tree.

Although these advanced search algorithms have achieved success in unimodal tasks, their application in the multimodal domain is still in its nascent stages. Recent research has begun to bridge this gap, VisuoThink [171] extends the foundational ToT algorithm into a forward-looking tree search algorithm. Meanwhile, Socratic-MCTS [172], vrest [173] and A star [174] have adapted MCTS to the vision-language domain. These methods model the visual question answering task as the construction of a reasoning tree composed of nodes representing (sub-question, sub-answer) tuples. A key innovation in this adaptation is the use of the model's own self-consistency score with respect to visual evidence as a reward signal to guide the tree search. This promotes the generation of more visually-grounded reasoning paths.

### 3.4 Dynamic Forensics During Reasoning

MLLMs typically employ a static, single-pass visual encoding mechanism when processing complex vision-language reasoning tasks, converting the image into a fixed representation before inference begins. However, a complex reasoning process demands that the model dynamically refocus its attention on the most critical visual evidence based on its

evolving inferential needs. The existing static mechanism restricts this dynamic interaction, making it prone to two critical flaws: first, the initial visual information tends to decay as text generation progresses. Secondly, the model may over-rely on its linguistic priors, generating hallucinations [175], [176] that contradict the visual facts.

The core solution is to establish a “think with image” [33] reasoning loop, thereby enabling the model to continuously revisit visual evidence during inference. As illustrated in Fig. 9, we illustrate the Think–Act–Observe reasoning loop. This section focuses on the central strategy for implementing such a loop: constructing interleaved vision-language Chains-of-Thought. Based on the source of the visual evidence used in this construction, the approaches can be classified into two main categories:

- **Vision-Language Alignment via Endogenous Visual Evidence Injection.** To enable a model to dynamically refocus on visual evidence during reasoning without using external tools, researchers have developed endogenous methods that leverage the model’s internal mechanisms. (Sec. 3.4.1).
- **Vision-Language Alignment via Exogenous Visual Evidence Injection.** Exogenous methods treat the Multimodal Large Language Model (MLLM) as an intelligent agent that actively gathers visual evidence by calling external tools, creating an “observe–act–decide” reasoning loop.(Sec. 3.4.2).

#### 3.4.1 Vision-Language Alignment via Endogenous Visual Evidence Injection

Endogenous methods aim to refocus on relevant visual information through the model’s internal attention mechanisms, without relying on external tools. Based on the coupling between the reasoning process and the visual encoder, these methods can be classified into two categories: those involving a **single forward pass** and those involving **multiple forward passes**. In Table 4, we provide an overview of the methods that takes this paradigm.

A **single forward pass** approach involves the encoder performing one global encoding of the visual input prior to inference. During the subsequent text generation and

TABLE 4  
An overview of representative methods about endogenous visual evidence injection.

Method	Venue	Granularity	Main Contribution
	<i>(a) Single forward pass.</i>		
CVC [177]	IJCAI'25	token	Proposing a mask-then-predict strategy that re-weights attention toward salient visual cues when needed.
ICoT [178]	CVPR'25	bbox	Selecting the most relevant regions via attention maps and re-encodes the corresponding crops.
MINT-CoT [147]	arXiv'25	token	Before each reasoning step, identify the most relevant visual tokens and re-inject them to guide generation.
Look-back [179]	arXiv'25	token	Invoking an internal look-back mechanism to revisit previously attended visual cues when inconsistency is detected.
	<i>(b) Multi forward passes.</i>		
CogCoM [145]	ICLR'25	bbox	Introducing chain-of-manipulations reasoning that iteratively localizes, acts on, and verifies regions.
DeepEyes [180]	arXiv'25	bbox	Using end-to-end RL to decide when to re-encode task-critical regions.
Pixel Reasoner [151]	arXiv'25	bbox	Training with SFT+RL to plan and execute visual operations and iteratively refine evidence.
SIFTthinker [150]	arXiv'25	bbox + depth	Simulating human visual search and augmenting with depth cues to strengthen spatial reasoning.
CMMCoT [181]	arXiv'25	bbox	Maintaining a memory bank to aggregate cross-image evidence for multi-image reasoning, revisiting key regions as needed.

reasoning phase, the model dynamically assigns higher attention weights to critical visual information based on the current reasoning context, thereby integrating visual evidence into the process. In contrast, a **multiple forward passes** approach allows the model to actively request a re-encoding of specific visual regions based on intermediate reasoning results. This enables it to obtain higher-resolution or more targeted features on an on-demand basis.

- **Single Forward Pass.** Initially, MM-COT [105] proposed a method to simulate refocusing. It first takes the image and the initial question as input to generate a Chain-of-Thought text containing key visual evidence, which is then concatenated back with the original input to derive the final answer. However, this approach suffers from a rigid interaction pattern and is computationally expensive. To address this issue, subsequent works have adopted strategies based on implicit and explicit attention guidance. **Implicit attention guidance** aims to redistribute attention through internal mechanisms without explicit visual cues; for instance, Look-back [179] achieves an implicit “re-look” by dynamically adjusting attention weights on key objects during text generation, while CVC [177] masks specific image regions to compel the model to backtrack and use other visual cues. **Explicit attention guidance**, in contrast, relies on clearly defined visual evidence. Point-RFT [182], for example, uses bounding box coordinates to direct the model’s attention to a specific area, whereas other works like Don’t Look Only Once [183], MINT-CoT [147] and ICOT [178] dynamically select the most relevant visual tokens from global features to aid the current reasoning step.

- **Multiple Forward Passes.** To obtain more reliable visual evidence, a line of research employs the strategy of multiple forward passes. Initial methods like Textcot [184] follow a sequential “describe-then-crop” paradigm, which generate a detailed image description to guide the cropping and reasoning over these image areas. However, this approach is prone to the loss of fine-grained visual details and suffers from an inflexible reasoning process. To overcome this, Studies such as CMMCoT [181] realize “think with image” paradigm by dynamically interleaving visual grounding

with the ongoing reasoning process. This is achieved by **adaptively** re-encoding key regions of interest (RoIs) at crucial steps and injecting these updated visual features into the subsequent reasoning chain. However, when to seek for more visual information is still an open question for community. Currently, some works like Sketchpad adopt in-context learning to follow prompts [185], works such as Chain-of-spot [186], PromViL [187], Vocot [188] adopt supervised fine-tuning to imitate expert paths, and works including Deepeyes [180], PixelReasoner [151], Active-O3 [189] and VLM-R<sup>3</sup> [190] adopt reinforcement learning to autonomously discover the optimal timing.

Further advancements build upon this paradigm, enhancing the quality and granularity of visual information acquired during reasoning. For instance, CogCoM [145] broadens the visual action space into a composable “chain of operations” like localization and labeling. Meanwhile, to improve spatial understanding, SIFTthinker [150], PixelThink [191] and LLaVA-Aurora [192] introduce depth information to enable joint object-level and spatial perception. This line of research has culminated in unified frameworks, such as the “RoI re-encoding and visual token re-sampling” proposed by Argus [193], which significantly enhances visual grounding while maintaining reasoning efficiency.

### 3.4.2 Vision-Language Alignment via Exogenous Visual Evidence Injection

Exogenous methods frame the MLLM as an intelligent agent that dynamically acquires visual evidence by calling external tools or interacting with an environment, creating an “observe-act-decide” reasoning loop. This paradigm has evolved from an early “plan-then-execute” model to a more flexible, interleaved Chain-of-Thought approach. In Table 5, we provide an overview of the methods that takes this paradigm.

Pioneering works such as Visual Programming [199], ViperGPT [195], HuggingGPT [200] and InternGPT [201] established a tool-assisted reasoning paradigm where the model performs a one-shot decomposition of a complex problem into a sequence of tool calls, often guided by in-

TABLE 5  
An overview of representative methods about exogenous visual evidence injection.

Method	Venue	Tools	Main Contribution
MM-ReAct [194]	(a) <i>In Context Learning.</i> arXiv'23	Visual experts	Proposing calling visual tools whenever a reasoning step requires additional evidence.
ViperGPT [195]	ICCV'23	GLIP/X-VLM/MiDaS	Introducing a program-as-agent paradigm that generates executable code to invoke visual tools.
CLOVA [196]	CVPR'24	OWL-ViT/BLIP/CLIP	Introducing learnable prompts to select and adapt tool usage during reasoning.
Visual Sketchpad [185]	NeurIPS'24	SAM/G-DINO/Matplotlib	Mimicing human problem solving by sketching while reasoning.
LLaVA-Plus [197]	(b) <i>Fine Tuning.</i> ECCV'24	OCR/SAM/BLIP2	Training on tool-grounded manipulation traces, improving the model's ability to use tools.
TACO [153]	ICLR 2025	DepthAnything/OCR/SAM	Invoking external tools when a step requires additional evidence or verification.
OpenThinking [198]	arXiv'25	Crop/Point/Zoomin	Proposing a unified API for visual operations to standardize and streamline tool invocation.

context learning (ICL) [202]. An executor then runs this pre-defined plan to generate a final answer. While this approach offers a clear planning path, its rigidity and lack of dynamic feedback make it brittle in complex, multi-step interactions and highly dependent on prompt engineering. To overcome this rigidity, subsequent research has shifted towards an adaptive, interleaved Chain-of-Thought model. In this paradigm, the model makes a decision at each reasoning step, selecting and executing a relevant visual operation and then using the immediate observation to inform its next action. The model acquires this step-by-step reasoning capability through either **in-context learning** or **fine-tuning**.

- **In-Context Learning.** Early works like ReAct [203] and MM-ReAct [194], along with later approaches such as Visual Sketchpad [185], primarily rely on the in-context learning (ICL) capabilities of powerful foundation models, using meticulously designed prompts to guide tool use. While this method is highly flexible, it places stringent demands on the model's instruction-following capabilities and depends heavily on high-quality prompt engineering. Consequently, its stability is often challenged in complex tasks. To improve the robustness of ICL, subsequent research like CLOVA [196] and VisRep [204] has introduced execution feedback and self-correction mechanisms.

- **Fine-Tuning.** To internalize tool-use capabilities as an intrinsic skill, a significant body of research has employed fine-tuning. For instance, works like LATTE [205], LLaVA-Plus [197], DWIM [206], Refocus [207] and VPD [208] use supervised fine-tuning on tool-use trajectory data to teach the model when and how to use tools. Building upon this, MLLM-Tool [209] expands the modality into audio. In pursuit of more optimal decision-making policies, VTool-R1 [210] and Visual-ARFT [211] introduce reinforcement learning, training the model to master the best timing and sequence for tool calls through trial and error. Furthermore, to enhance the generality and usability of this paradigm, researches such as OpenThinking [198] has encapsulated diverse visual tools into a unified API, allowing the model to dispatch them on-demand during inference.

In summary, this section has systematically reviewed the cutting-edge methodologies for enhancing the interactive reasoning capabilities of MLLMs, structured along our

proposed "From Perception to Cognition" framework. At the **perceptual level**, the research has progressed along two major trends: one involves adopting dynamic strategies that integrate multiple expert encoders to capture richer, multi-granularity features; the other focuses on unifying image generation and understanding tasks, leveraging the powerful representational capabilities of generative models to enhance the perceptual foundation.

At the **cognitive level**, the focus has shifted from initial reliance on prompt-engineered, single-path imitation learning to a more systematic, multi-paradigm fusion. This includes supervised fine-tuning on visually-grounded Chain-of-Thought data, employing preference learning (e.g., DPO/GRPO) to select among multiple candidate paths, and utilizing tree-based search at inference time to explore a more optimal solution space. This convergence of methods aims to endow models with greater autonomy in problem decomposition and dynamic verification.

This synergistic development at both the perceptual and cognitive levels has collectively driven significant advancements in MLLM capabilities. However, to objectively measure the effectiveness of these methods and to reveal their strengths and limitations in specific application scenarios, a systematic evaluation framework is essential. In the next chapter, we will delve into the key benchmarks and applications used to assess these advanced models.

## 4 APPLICATIONS AND BENCHMARK

### 4.1 Scientific Problem Solving

Early multimodal large models primarily focused on fundamental visual-language alignment. In the domain of visual-textual reasoning, pioneering benchmarks such as VQA [212] and its successor, VQA-v2 [213], trained models to answer basic questions like "What is in the image?" Subsequently, GQA [214] enabled preliminary visual reasoning by introducing compositional questions. However, these models operated predominantly at the perceptual level, concentrating on object recognition. While subsequent developments introduced benchmarks such as VCR [215], which requires models to provide rationales; OK-VQA [216] and A-OKVQA, which incorporate external knowledge; and LoRA [217], which systematically evaluates logical reasoning

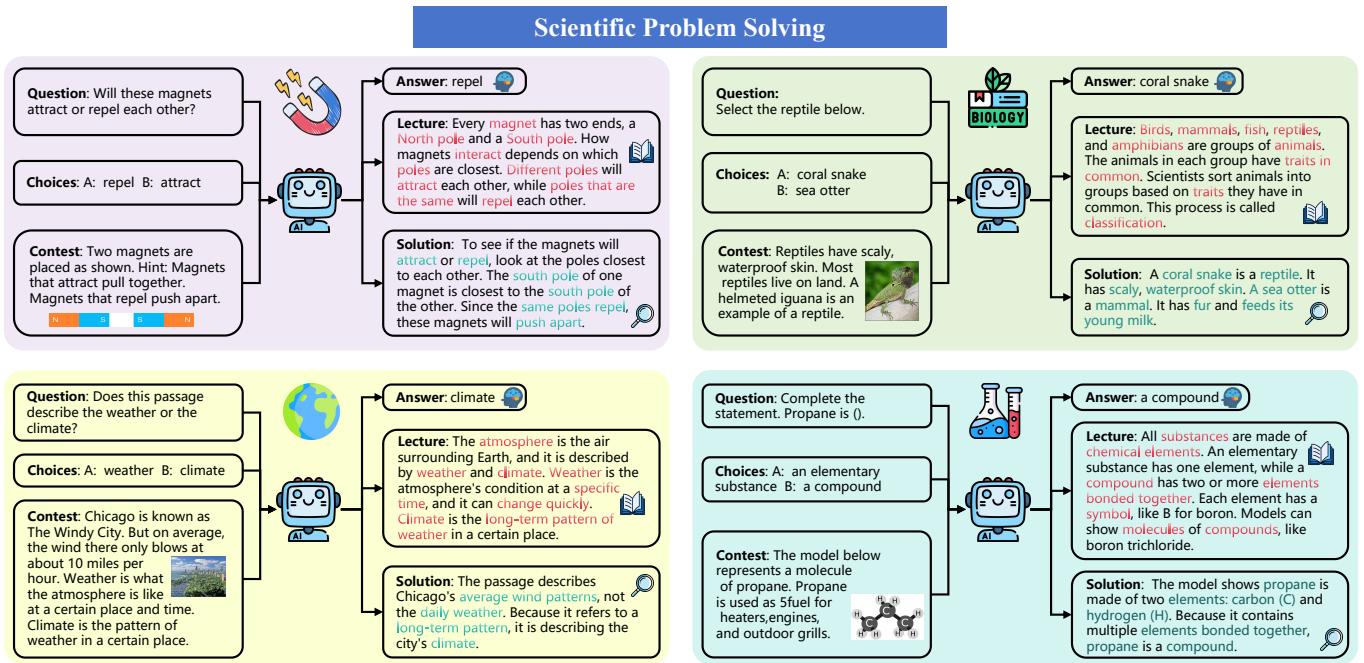


Fig. 10. Scientific Problem Solving Dataset Examples, featuring questions in physics, biology, geography, and chemistry.

capabilities, they still fell short of addressing the core challenge in scientific problem-solving: abstract symbolic reasoning [218]. Although several specialized evaluation benchmarks have been designed, including RAVEN [219] for analogical reasoning, AlgoPuzzleVQA [220] for algorithmic reasoning, the Spatial Commonsense Benchmark [221] for spatial awareness, and Fig-QA [222] for figurative language understanding, these benchmarks remain insufficient for adequately measuring the rigorous, multi-step, and deterministic deductive reasoning required in the scientific domain [223].

To address this challenge, ScienceQA [224] was one of the first and most complex benchmarks created specifically for scientific problem-solving. Its dataset comprises over 20,000 multimodal questions from grade 3-12 science curricula, covering fields such as natural sciences, social sciences, and language arts. A distinguishing feature of ScienceQA is that each question is annotated with a complete chain of thought [225]. This design enables not only the evaluation of the final answer's correctness but also the scrutiny of the logical coherence of the model's reasoning process. In Fig. 10, we provide several examples drawn from scientific problem-solving datasets. Another critical benchmark is MathVista [226], which systematically integrates 28 existing visual-mathematics datasets to create a comprehensive evaluation suite. This suite covers seven core reasoning types, including algebra, geometry, and statistics [218], and features a dedicated section of IQTest [226] puzzles.

For a more comprehensive assessment, MathVerse [227] and MATH-V [228] provide high-quality problems sourced from actual mathematics competitions. Concurrently, to extend evaluation into broader scientific disciplines, MDK12-Bench [229] utilizes authentic K-12 examination questions, and SCI-Reason [230] employs multi-panel illustrations from scientific papers. To assess model robustness, MATHREAL [231] constructs its dataset by introducing real-world noise, while MATHCHECK [232] and UGMathBench [233] test

model stability through diversified tasks. Similar benchmarks include EMMA [234], for evaluating multidisciplinary reasoning; MV-MATH [235], for handling multiple visual inputs; RMath [236], with a focus on logical reasoning; and MMMU [237], which concentrates on multimodal mathematical understanding. These benchmarks are built upon foundational datasets like GSM8k [238] and MATH [239]. To prevent performance saturation of leading models on existing benchmarks, R-Bench [240] and MR-MATH [241] have begun to incorporate problems from the graduate level and even from contemporary mathematical research.

We have summarized the performance of mainstream multimodal models on several scientific problem-solving benchmarks, namely MathVista, MathVerse, MATH-V, MV-MATH, and MMMU, as presented in Table 6. The evaluation results indicate that among proprietary models, Gemini 2.5 Pro demonstrated exceptional, state-of-the-art performance, achieving the highest scores across all five benchmarks. This indicates its superior capabilities in handling cognitive tasks such as complex symbolic operations, geometric-spatial imagination, and multi-step logical deduction. This superior performance can be attributed to its advanced internal reasoning architecture, which employs mechanisms analogous to a chain of thought to conduct deep, multi-path exploration and verification of problems [258], thereby significantly enhancing its solution accuracy for complex mathematical problems. Among open-source models, the InternVL3 series exhibited outstanding performance, particularly the version enhanced with VisualPRM-Bos [259], which improved scores by 4 to 10 percentage points compared to the baseline version. The core advantage of this technology lies in its adoption of process supervision, which replaces the conventional outcome supervision that focuses only on the final result. Specifically, during the reasoning phase, the model first employs a Best-of-N sampling strategy to generate multiple, logically diverse candidate reasoning

TABLE 6

Performance comparison on scientific problem-solving benchmarks. The best-performing model's scores are highlighted in bold.

Models	Math Vista (Acc ↑)	MathVerse (Acc ↑)	MATH-V (Acc ↑)	MV-MATH (Acc ↑)	MMMU (Acc ↑)
Human	60.3	-	68.82	76.5	88.6
Random Choice	17.9	12.4	7.17	-	22.1
<i>Proprietary Models</i>					
Gemini 2.5 Pro [242]	<b>84.6</b>	<b>84.6</b>	67.3	<b>73.3</b>	<b>82.0</b>
Gemini 2.5 Flash [242]	81.2	-	-	-	79.7
Gemini 2.0 Pro [243]	-	67.3	48.1	-	69.9
Gemini 2.0 Flash [243]	72.2	47.8	43.6	-	72.6
Gemini 1.5 Pro [244]	68.3	68.1	-	29.1	65.8
GPT-5 [245]	82.7	-	-	-	-
GPT-4o [25]	63.8	40.6	31.2	32.1	70.7
GPT-4 Turbo [246]	66.8	66.8	46.7	-	56.8
GPT-4V [247]	-	-	-	-	56.0
Grok-3 [248]	-	-	-	-	78.0
o1 [249]	73.9	-	-	-	78.2
Claude 3.5 Sonnet [250]	67.7	46.7	41.9	33.9	75.0
Qwen-VL-Max [251]	-	-	-	42.4	75.0
Kimi-k1.5 [252]	74.9	-	38.6	-	70
<i>Open-Source Models</i>					
InternVL3 (78B) [253]	75.1	48.2	34.2	-	72.2
InternVL3 (38B) [253]	75.1	<b>44.4</b>	37.2	-	70.1
InternVL3 (14B) [253]	71.6	39.8	29.3	-	67.1
InternVL3 (8B) [253]	71.6	39.8	29.3	24.2	62.7
QVQ-72B-Preview [254]	71.4	-	35.9	29.3	70.3
Qwen2.5-VL (72B) [23]	74.8	57.6	38.1	-	70.2
Qwen2.5-VL (8B) [23]	67.8	41.1	25.4	-	55.0
LLaVA-OneVision (72B) [255]	67.1	27.2	25.3	-	56.8
LLaVA-OneVision (7B) [255]	58.6	19.3	18.3	-	48.8
Llama 3.2 (90B) [256]	57.3	-	-	-	60.3
Ovis2 (34B) [257]	76.1	50.1	31.9	-	-
Ovis2 (16B) [257]	-	45.8	30.1	-	-

paths. Subsequently, a fine-tuned Visual Process Reward Model (Visual PRM) functions as an external evaluator to perform a granular assessment and scoring of each step within every candidate path. The system ultimately selects the path with the highest cumulative reward, representing the most validated reasoning process, as the final output. This synergistic "generate-and-verify" [225] mechanism ensures that the correctness of the answer is founded upon a logically rigorous and procedurally reliable process, rather than being an incidental outcome. This approach significantly enhances the robustness and accuracy of the model's reasoning.

Collectively, the performance of these models is strongly correlated with their capabilities for robust chain-of-thought generation, multi-path exploration and verification, and self-evaluation and correction during the reasoning process. Future advancements will need to focus on enhancing the cognitive intelligence of these models and their capacity for knowledge integration and innovation when confronted with novel, open-ended problems [218]. The key challenge is to transition from solving closed-domain textbook problems to addressing open-ended, research-level questions [260], while ensuring that their reasoning processes are not only correct in their outcomes but also logically aligned with human cognitive patterns.

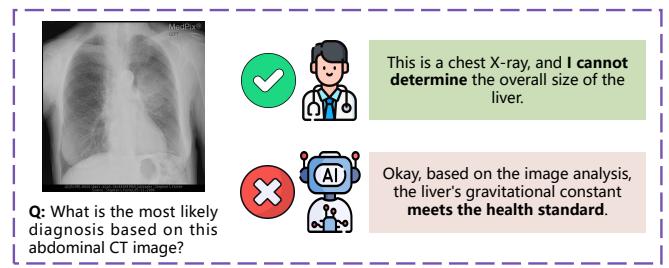


Fig. 11. An example of a FAKE question. Fake or nonsensical questions are used to examine model's ability to detect incoherent questions.

## 4.2 Medical Diagnosis

Medical imaging, which directly reflects the state of the human body, is a critical component of clinical decision-making. In the medical field, even minor errors can lead to severe consequences. Consequently, the focus of evaluation has rapidly shifted toward reliability, with a particular emphasis on combating "hallucinations." Early medical VQA benchmarks, such as VQA-RAD [261], laid the groundwork for the field. Constructed by clinicians using authentic radiological images, its questions and answers are typically concise and focus on the identification of anatomical structures and abnormal findings. However, with the advancement of generative models, the phenomenon of fabricating plausible but incorrect information has emerged as the most significant safety concern in medical AI. To address this, HALT-MedVQA [262] was specifically designed. In addition to standard medical question-answer pairs, its construction involved the deliberate introduction of numerous nonsensical or factually contradictory probe questions, in Fig. 11, we provide an example of a probe question. The primary objective of this benchmark is not to measure how many questions a model can answer correctly, but to test its ability to recognize and refuse to answer inappropriate queries, akin to the rigorous standards of a medical professional.

The success of VQA-RAD spurred the creation of a series of benchmarks targeting different medical specialties. For example, PathVQA [263] focuses on pathology images, Kvasir-VQA [264] on gastrointestinal diagnosis, EndoVis-17/18-VQLA [265] on robotic surgery scenarios, and MicroVQA [266] on expert-level scientific reasoning for biological microscopy images. In parallel, benchmarks such as Med-VQA [267], SLAKE [268], and PMC-VQA [269] have been developed to construct larger-scale, more comprehensive general medical question-answering datasets covering a wider range of modalities. The intense focus on the core issue of "hallucinations" has also led to the development of systematic evaluation frameworks beyond HALT-MedVQA, including MedHallBench [270] and MedHallMark [271].

We have summarized the performance of various open-source and proprietary large models in the field of medical diagnosis, as shown in Table 7. The results indicate that while the most advanced models have approached or even surpassed human-level performance on certain tasks like VQA-RAD, a significant gap remains on benchmarks requiring sophisticated pathological understanding and knowledge-based reasoning, such as PathVQA, where they lag behind human experts (85.2%). Among all models, Med-PaLM M

TABLE 7

Performance comparison on medical diagnosis benchmarks. The best-performing model's scores are highlighted in bold.

Models	VQA-RAD SLAKE	Path-VQA	Med-VQA	PMC-VQA
	(Acc ↑)	(Acc ↑)	(Acc ↑)	(Acc ↑)
Human	77.3	93.4	85.2	-
<i>Proprietary Models</i>				
Gemini 1.0 Pro [275]	73.4	79.5	56.6	62.0
GPT-4V	86.6	85.8	66.8	76.5
Med-PaLM M [272]	<b>90.0</b>	<b>90.5</b>	<b>75.0</b>	<b>81.3</b>
<i>Open-Source Models</i>				
CogVLM (17B) [276]	78.0	83.9	60.1	63.9
CogVLM (Chat) [276]	79.8	84.1	61.3	-
InstructBLIP (13B) [99]	58.3	76.2	45.4	-
LLaVA-Med (13B) [76]	72.5	82.1	54.0	50.2
LLaVA-Med (7B) [76]	68.0	79.5	49.6	50.2
Med-Flamingo (80B) [277]	81.5	86.8	70.3	-
MedViNT (13B) [269]	75.4	83.1	58.2	-
MedViNT (7B) [269]	71.3	81.0	53.6	-
MiniGPT-v2 (7B) [278]	63.2	77.1	45.4	47.7
Qwen-VL-Chat (9B) [279]	61.2	76.5	49.2	48.9
Qwen-VL-Chat (1.8B) [279]	53.4	69.8	41.6	-

[272], a proprietary model specifically designed for the medical domain, demonstrated overwhelming superiority, achieving the highest scores across all five benchmarks. Its outperformance is attributable to a meticulously designed evaluation and training framework that spans multiple medical tasks. Through instruction fine-tuning, the model has learned to align visual information with specific medical instructions and knowledge. It does not merely recognize content within an image; it learns to interpret the image in the context of a specific question, providing answers that are consistent with medical logic. This represents a crucial leap from perception to cognition. The significantly lower accuracy of all models on the PathVQA dataset compared to human experts reveals the limitations of current AI in highly specialized cognitive tasks. PathVQA's focus on pathology images requires an exceptionally detailed and professional understanding of cellular morphology, tissue architecture, and staining characteristics [273]. This level of depth and nuance is difficult for general-purpose or even general medical models to achieve. To answer such questions, a model must establish a tight connection between fine-grained visual features and abstract medical concepts [3]. Human experts develop this robust cognitive link through years of dedicated training, a process that is exceedingly difficult to replicate for a model learning from a limited dataset [274]. Therefore, a critical challenge for current large models is to enhance their knowledge integration and cognitive reasoning capabilities within specialized professional domains [273]. This represents the most formidable step in transitioning from current perceptual intelligence to a truly human-like cognitive intelligence.

### 4.3 Diagram Understanding

Diagram understanding poses unique challenges for multimodal models that extend beyond conventional visual perception. This task requires models to integrate visual

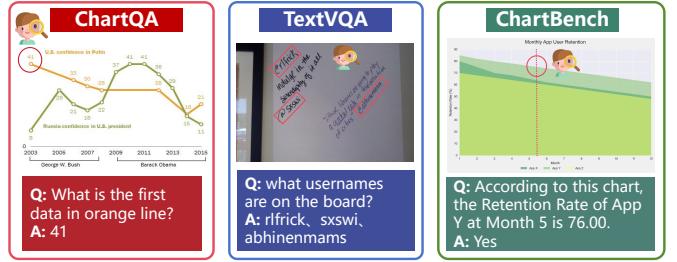


Fig. 12. Examples from the ChartQA, TextVQA, and ChartBench Datasets.

perception, text comprehension, and numerical logical reasoning to extract and infer abstract data from structured visual information. Existing general-purpose multimodal benchmarks, such as the comprehensive evaluation standards of MMBench [280], MME [281], MM-Star [282], and MM-Vet [283], often lack in-depth assessment of such capabilities.

To better evaluate this ability, a specialized task benchmark for chart question answering was introduced: ChartQA [284]. Its construction combines two methodologies. One portion of the questions is formulated by human experts based on charts, ensuring the naturalness and complexity of the problems. The other portion is generated through semi-automated templates, which significantly expands the dataset's scale and coverage. The human-annotated subset, ChartQA-Human [284], serves as the test set to determine if a model has genuinely acquired human-like chart interpretation skills.

However, a limitation of ChartQA is that many of its charts are accompanied by underlying structured data tables, which allows models to potentially exploit a shortcut to answer questions [285]. Prior to ChartQA, PlotQA [286] represented an early attempt to construct a large-scale chart question-answering benchmark, but its heavy reliance on synthetic charts limited its ability to capture the complexity of real-world scenarios. To resolve the "tabular shortcut" problem, the subsequent ChartBench [287] was developed by removing all tabular data. This change compels models to reason exclusively from visual elements, such as bar heights and line slopes, in a manner akin to human interpretation [288]. The development in this area is built upon a series of foundational works, evolving from early scene text recognition (e.g., TextVQA [289]) and document understanding (e.g., DocVQA [290]) toward more sophisticated evaluation paradigms. For example, ChartMind [291] extends the evaluation to more complex real-world scenarios, proposing open-ended tasks like trend analysis and data insight summarization. Meanwhile, ChartMimic [292] introduces a more challenging generative task, requiring the model to generate code that can reproduce a chart based on an example image and instructions. In Fig. 12, we provide examples from ChartQA, TextVQA, and ChartBench to highlight the distinctions among the datasets.

In Table 8, we summarize the performance of a range of leading open-source and proprietary large models on six mainstream diagram understanding benchmarks. These tasks go beyond basic visual perception and impose rigorous demands on the models' cognitive reasoning abilities, including chart parsing, data extraction, spatial relationship understanding, and mathematical computation. The data reveal

TABLE 8

Performance comparison on diagram understanding benchmarks. The best-performing model's scores are highlighted in bold.

Models	ChartQA	PlotQA	InfographicVQA	DocVQA	TextVQA	TabMWP (Acc ↑)
<i>Proprietary Models</i>						
Gemini 1.5 Pro	87.2	73.1	<b>81.0</b>	93.1	78.7	96.9
GPT-4o	84.1	<b>74.5</b>	60.1	91.5	82.3	<b>97.4</b>
GPT-4V	78.4	68.9	55.2	88.6	78.3	95.3
Claude 3 Opus [296]	82.5	72.8	57.5	90.2	80.7	96.5
Qwen-VL-Max	79.8	69.3	-	93.1	79.5	95.8
<i>Open-Source Models</i>						
InternVL-Chat-V1.5 (34B) [297]	83.8	71.5	54.3	88.5	78.9	-
InternVL-Chat-V1.5 (8B) [297]	79.1	68.2	51.2	85.1	75.6	-
Qwen2.5-VL (72B)	<b>89.5</b>	-	-	<b>96.4</b>	-	-
Qwen2.5-VL (7B) [23]	87.3	-	-	95.7	<b>84.9</b>	-
Qwen-VL-Chat (9B)	75.8	65.4	-	83.5	74.3	92.1
Qwen-VL-Chat (1.8B)	68.2	59.7	-	78.9	69.1	85.6
LLaVA-NeXT (34B) [298]	77.2	66.8	51.5	84.8	76.5	91.2
LLaVA-NeXT (7B) [298]	73.5	62.1	48.3	81.2	72.8	88.4
DeepSeek-VL (7B) [299]	76.3	66.1	-	84.1	75.0	-

that proprietary models, led by GPT-4o and Gemini 1.5 Pro, demonstrate powerful and well-balanced comprehensive capabilities, setting high performance benchmarks across most tasks. Notably, GPT-4o leads on the TabMWP [293] task, which emphasizes mathematical and logical reasoning, with an accuracy of 97.4%, a strength attributed to its powerful general-purpose reasoning abilities. Concurrently, Gemini 1.5 Pro performs best on InfographicVQA [294], which requires parsing complex and unstructured layouts, underscoring its advantages in processing information-dense visual content. This feature is closely linked to its Mixture-of-Experts architecture, which can process vast contexts. Among open-source models, Qwen2.5-VL (72B) achieves the highest accuracy on two key benchmarks, ChartQA and DocVQA, where its performance surpasses not only all other open-source models but also the top proprietary models. This model employs a native dynamic-resolution Vision Transformer and advanced omni-document parsing techniques. This allows it to process high-resolution inputs without downsampling, enabling it to accurately capture and understand fine-grained text, layout structures, and spatial relationships within diagrams, thus achieving a profound leap from perception to cognition.

Overall, models in the domain of diagram understanding are evolving from 'information processors' to 'cognitive agents' [295]. They are capable not only of understanding visual content but also of reasoning, planning, and executing complex tasks that interact with the digital world based on that content. Future models will be expected not only to "understand" diagrams but also to perform deep reasoning, dynamic interaction, and cross-domain knowledge integration based on them, which will allow them to play a central role in a broader range of real-world applications.

#### 4.4 Video Understanding

Video understanding extends visual-textual reasoning from static images to dynamic sequences, introducing complex dimensions such as time, variation, and causality [300]. This requires a model not only to recognize content but also to elucidate the underlying reasoning logic. Early research established a foundation for explainability on static images

through benchmarks like VQA-E [301] and introduced the "Visual Chain of Thought" (Visual CoT) methodology. However, the dynamic nature of video necessitates more complex, graph-structured reasoning capabilities.

To address the causal complexity inherent in video, researchers have constructed specialized benchmarks. Representative works include CausalVQA [302], which deeply investigates causal relationships through five distinct question types: descriptive, counterfactual, and predictive, among others. Another is VCRBench [303], which requires models to reorder shuffled video clips into a logical sequence, directly testing their ability to model long-range causal dependencies. Building on this, CausalStep [304] enhances the rigor of evaluation by compelling models to perform step-wise reasoning. However, the powerful generative capabilities of modern models are accompanied by the problem of "hallucinations" in the video domain. Consequently, a series of specialized diagnostic benchmarks has emerged. Vid-Halluc [305] focuses on evaluating temporal hallucinations, while HAVEN [306] and VideoHallucer [307] provide more comprehensive frameworks for assessing and mitigating both intrinsic and extrinsic hallucinations in video understanding. Additionally, specialized benchmarks such as AVHBench [308] for audio-visual content and Hallu-PI [309] for handling input perturbations have been developed.

We have summarized the performance of various open-source and proprietary large models on several cutting-edge video understanding benchmarks, as shown in Table 9. The results show that open-source models demonstrate exceptional performance on specific high-difficulty benchmarks, whereas the evaluation data for proprietary models are often sparse and opaque. Furthermore, a significant gap persists between all models and human performance on complex cognitive reasoning tasks. Among proprietary models, Gemini 1.5 Pro, with its massive context window supporting up to ten million tokens and its Mixture-of-Experts (MoE) architecture, exhibits near-perfect performance in long-video question answering and key information retrieval. Among open-source models, Qwen-2.5-VL and STORM each excel in different areas, with their superior performance demonstrated on benchmarks designed for distinct cognitive dimensions. The MLVU [310] benchmark, which focuses on long-video understanding with an average video duration of 15 minutes, includes diverse genres like movies and vlogs to evaluate a model's integrated ability to process both holistic information and local details. Qwen-2.5-VL stands out on this benchmark due to its unique architectural innovations. The model employs an "Absolute Time Encoding" technique that directly aligns the IDs of its Multimodal Rotational Position Encoding (MRoPE) with video timestamps. This enables second-level precision for event localization without introducing additional computational layers, a crucial feature for identifying specific details in extended videos. On the other hand, the MVBench [311] benchmark is specifically designed to evaluate temporal reasoning, comprising 20 dynamic tasks that cannot be solved with single-frame analysis, thus testing a model's temporal skills from perception to cognition. Fig. 13 further illustrates examples of key task categories, such as 'Action Antonym', 'Action Localization', and 'Scene Transition'. STORM achieves leading performance on this benchmark, with its core advantage lying in its novel architec-

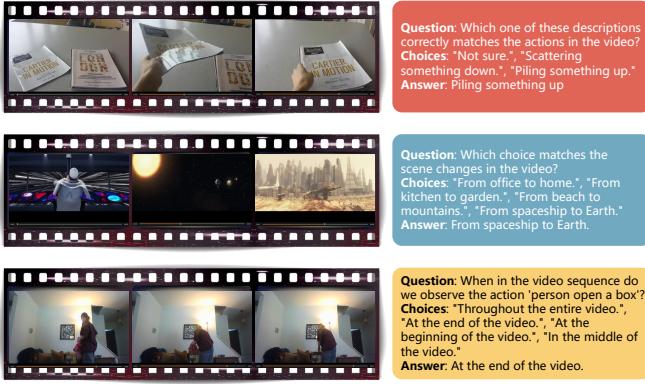


Fig. 13. Examples of Key Tasks for Evaluating Video Understanding: Action Antonym, Action Localization, and Scene Transition.

ture [312]. It integrates a dedicated temporal encoder, based on a Mamba state-space model, between the image encoder and the large language model. This module explicitly models inter-frame dynamic relationships, effectively preserving critical temporal information throughout the video sequence. This allows it to accurately capture and reason about complex temporal variations, perfectly aligning with the evaluation focus of MVbench. In parallel, VideoMME [313] serves as a large-scale, multi-dimensional evaluation for video understanding. Although it primarily focuses on general-purpose tasks, its cross-modal question answering and fine-grained description components also involve challenges related to causality and hallucination. Consequently, it is often used in practice as a key supplementary benchmark for assessing a model's overall robustness and trustworthiness.

The primary challenges currently facing the field of video understanding are the fragmentation and opacity of its evaluation ecosystem. Models are often tested on disparate benchmarks, making fair comparisons difficult. Concurrently, issues of data contamination and "shortcut" learning are prevalent, leading to inflated performance scores. Future research must shift from processing clean, curated data to handling real-world noise and from closed-ended question answering to open-ended problems that require knowledge integration. This necessitates an urgent, collaborative effort from the academic community to construct a unified, transparent evaluation framework capable of measuring high-level cognitive abilities.

#### 4.5 Sentiment Analysis

Although current multimodal large models can readily identify static facial expressions, a substantial gap remains in their ability to comprehend the complex world of human social and emotional dynamics. Existing benchmarks for knowledge retrieval, such as Infoseek [318], or for factual reasoning, like VQA, are not relevant to sentiment analysis. Meanwhile, commonsense reasoning benchmarks, such as SocialIQA [319], are predominantly text-based and lack the authentic interactivity of multimodal social scenarios [320]. Evaluating a model's "Emotional Quotient" (EQ) requires moving beyond simple expression classification and into the realm of dynamic, multi-party, and realistic social interactions.

To advance and test the "EQ" of models, researchers have proposed a range of specialized benchmarks. The milestone

TABLE 9  
 Performance comparison on video understanding benchmarks. The best-performing model's scores are highlighted in bold.

Models	CausalVQA	VCRBench	MLVU	MVBench	Video-MME
	(Acc ↑)	(Acc ↑)	(M-Avg ↑)	(Score ↑)	(w/o subs)
Human	84.8	96.4	-	-	-
<i>Proprietary Models</i>					
Gemini 1.5 Pro	-	<b>48.2</b>	-	-	<b>75.0</b>
GPT-4o	<b>51.0</b>	29.0	54.9	64.6	71.9
<i>Open-Source Models</i>					
STORM [314]	-	-	72.9	<b>70.6</b>	-
InternVL2.5 (78B) [315]	47.5	14.5	59.9	-	-
Qwen2.5-VL (72B)	-	29.0	<b>74.6</b>	70.4	73.3
Qwen2.5-VL (7B)	49.1	7.1	70.2	69.6	65.1
LLaVA-Video-72B [316]	-	5.2	-	64.1	-
Video-LLaVA (7B) [317]	-	-	47.3	58.6	-



Fig. 14. An Illustrative Sentiment Analysis Example from the MELD Dataset (Derived from the TV show *Friends*).

MELD dataset, derived from the American television series *Friends*, utilizes multi-party dialogue videos to evaluate a model's understanding of context, character interactions, and emotional dynamics, Fig. 14 shows a dialogue scene. Subsequent benchmarks have focused on more nuanced cognitive challenges. For example, CA-MER [321] concentrates on scenes of emotional conflict, such as a character smiling while expressing sad sentiments, to test a model's reasoning capabilities in situations of cognitive dissonance. Building on this, HumanVBench [322] systematically evaluates the alignment between internal emotions and external expressions, while HumaniBench [323] is the first benchmark designed around human-centric AI principles like fairness and empathy. To probe deeper levels of social cognition, Social-IQ [324] assesses the understanding of social dynamics, and GD-VCR [325] and CVQA [326] are dedicated to addressing the problem of cultural bias. The trend in this field is shifting from classification toward more fine-grained tasks such as tracking, with benchmarks like VEATIC [327], and generative understanding, as seen in MER-UniBench [328]. Additionally, other benchmarks include MTMEUR [329], which focuses on multi-turn dialogue reasoning; Multi-HM [330], designed for specific scenarios; and the MERR Dataset [331], which provides large-scale data for training emotion recognition and reasoning.

We have summarized the performance of various models on multimodal sentiment analysis benchmarks. As shown in Table 10, the proprietary model Gemini 2.5 Pro achieved a leading score on the MME-EMOTION [334], leveraging its built-in "thinking" mechanism. However, its success rate is not as high as in other domains. Furthermore, the evaluation data for two key benchmarks, MELD [335] and HumanVBench, are extremely sparse [336]. Only a few

TABLE 10

Performance comparison on sentiment analysis benchmarks. The best-performing model's scores are highlighted in bold.

Models	MME-EMOTION (Rec ↑)	MELD (WF1 ↑)	HumanVBench (Avg Emo. Perc. ↑)
<i>Proprietary Models</i>			
Gemini 2.5 Pro	<b>39.3</b>	-	-
Gemini 1.5 Pro	32.8	-	-
GPT-4o	37.8	-	-
<i>Open-Source General Purpose Models</i>			
InternVL 2.5 (20B) [315]	29.2	<b>45.0</b>	-
InternVL-Chat-V1.5 (26B) [297]	20.8	-	<b>36.0</b>
LLaVA-OneVision (72B)	37.9	-	-
Qwen2.5-VL (72B)	31.3	39.1	25.8
Qwen2.5-VL (7B)	28.4	-	-
Qwen2-VL (7B) [332]	31.1	-	35.4
VideoLLaMA2 (7B) [333]	29.8	32.7	25.8

open-source models, including InternVL, Qwen, and VideoLLaMA2, have reported results, while no data are available for any proprietary models or most other open-source models. This comparison highlights a systemic evaluation dilemma within the field. Because proprietary model developers tend to release results on their own comprehensive, private benchmarks, a global performance comparison deficit is created [337]. This makes it difficult for researchers to fairly and comprehensively assess the capabilities of different models on specific cognitive dimensions under a unified standard. This data gap is particularly pronounced between the open-source and proprietary ecosystems, and it represents a formidable obstacle to measuring the field's true progress from perceptual intelligence toward cognitive intelligence.

## 5 FUTURE DIRECTION

In response to the present challenges, some future research directions can be inferred to build the next generation of Multimodal Large Language Models (MLLMs) capable of truly bridging the gap from perception to cognition.

### 5.1 Unified Vision Encoder

Despite rapid progress, current visual encoders of MLLMs often fail to comprehensively capture task-relevant visual information, leaving the evidence available to language reasoning incomplete. Building on this observation, recent work explores unified visual encoders that provide multi-granular, more comprehensive representations, integrating understanding and generation as well as multiple visual modalities within a single framework. For example, ATO-KEN [338] encodes images, videos, and 3D assets into a shared latent space, aiming to unify both tasks and modalities in a single framework. By fusing CLIP-level semantics with unified autoregressive training, TokLIP [339] equips visual tokens with high-level semantic understanding and enhances their low-level generative fidelity. However, this unification remains incomplete: a persistent gap separates understanding and generation, and a true integration across visual modalities has yet to be achieved. Therefore, developing a truly unified, multi-granular visual encoder,

with strong alignment and efficiency, remains a valuable direction for future work.

### 5.2 Latent Reasoning

Recently, a line of research has emerged to explore direct intervention within the latent space to guide the reasoning process of vision-language models. Different from traditional approaches that operate on input or output layers, these methods act directly on the model's latent representations, enabling more flexible and fine-grained control. Several key strategies exemplify this approach. For example, Multimodal Chain of Continuous Thought [340] introduces a method for continuous reasoning by iterating on "thought vectors" within the latent space, which enhances the alignment and fusion of cross-modal information. Concurrently, VTI [341] stabilizes visual and textual features to alleviate hallucinations by injecting corrective directions during the inference stage. Similarly, jiang *et al.* [342] leverages orthogonalization to directly nullify directions in the latent space associated with spurious concepts.

The synthesis of these distinct approaches presents a promising path for future work: exploring how to simultaneously achieve continuous reasoning, enhanced robustness, and hallucination suppression. The ultimate goal is to construct multimodal reasoning frameworks that are significantly more controllable and interpretable. The ultimate goal is to construct multimodal reasoning frameworks that are significantly more controllable and interpretable. By directly shaping the model's internal cognitive processes, this line of research promises to build a more robust bridge between raw visual perception and reasoned linguistic output, fostering a more grounded and coherent understanding of the visual world for more reliable interactive reasoning.

### 5.3 Generative Reasoning

This paradigm externalizes the model's implicit reasoning process into explicit visual entities as perception input for subsequent steps, proving highly valuable in domains such as robotic planning, visual puzzle solving, and dynamic simulations. Works like Chameleon [343] provide the foundational architecture for this paradigm by achieving mixed-modal understanding and generation of text and images, allowing for interleaved output in any sequence. Subsequent research, such as Visual Planning [344] aims to ground the planning and reasoning processes in visual domain. More recently methods like MVoT [345], Mind's Eye of LLMs [346] and ViLaSR [347] prompt the model to generate and iteratively update visual scratchpads when solving complex problems, and then use these visual "thought traces" to drive subsequent reasoning.

However, despite these methods have prospective future, they still face some challenges: The generated intermediate images can be inaccurate or contain hallucinations, failing to match the original source. And more importantly, the curation of suitable training data presents a significant challenge. To address these issues, future research can explore how to enhance the generation quality and how to reduce the dependency on manually curated datasets.

## 5.4 Tool-Augmented Reasoning

As we discuss in Sec. 3.4.2, representative works such as PixelReasoner [151], and OpenThinking [198] demonstrate significant advancements in both training paradigms and the richness of the tool ecosystem. However, there still exist some problems:

- There is often a consistency gap between visual cues and the reasoning process, leading to ungrounded conclusions. While methods like GThinker [348] introduce verification steps to check visual cues during reasoning, this often comes at the cost of increased inference time, creating an efficiency-accuracy trade-off.
- Current models often generate linear reasoning paths, which limits their ability to solve complex, multi-step problems that require deeper problem composition or exploration of multiple possibilities.

To address these problems, future research can focus on optimizing both the structure of the reasoning path and the timing of tool use. To create more sophisticated reasoning trajectories, tree-based algorithms like MCTS [128] can be explored. Simultaneously, the challenge of when to use a tool can be addressed by designing adaptive mechanisms. Such mechanisms should aim to ensure the correctness and consistency of visual cues while balancing the trade-off between verification accuracy and inference speed.

## 5.5 Cross-Image Relation Reasoning

Cross-image relation reasoning refers to the advanced capability of reasoning across multiple images. This requires a model to comprehend the logical or sequential relationships between events depicted in a series of images in order to draw a final conclusion. However, the vast majority of methods [94], [95], [180] discussed in this survey concentrate on single-image reasoning. Only a few works focus on the multi-image problems. For instance, CmmCoT [181] incorporates a memory bank to retrieve relevant information from associated images during inference time. Focus-Centric Visual Chain [349] explicitly models the inter-image relationships, while Mantis [350] introduces an interleaved instruction tuning. These works point out an open question for the future: **how to reduce memory loss of image evidence at inference time and enable more flexible mining of inter-image relationships?** Addressing this question is a critical step toward endowing MLLMs with a form of contextual memory, allowing them to perceive and reason about the world as a continuous stream of interconnected events rather than a series of isolated snapshots.

## 5.6 Real-Word Cognitive Evaluation

A major limitation of current evaluation systems is their reliance on clean data and closed-ended question-answering formats, which creates a significant gap between them and the dynamic real world, as well as advanced human cognitive activities. **On the one hand**, existing benchmarks are mostly built on rigorously cleaned and precisely annotated data, whereas real-world environments are filled with multi-source interference such as visual occlusions, background noise, and lighting variations. **On the other hand**, current tasks generally remain at the perceptual level and fail to adequately

address higher-level cognitive reasoning. For instance, even in the high-stakes application of medical diagnosis, models still perform significantly worse than human experts on the PathVQA [263] benchmark.

Therefore, future cognitive evaluation must pivot from closed environments to the complex challenges of the real world. **The benchmarks must transcend basic perception to focus on higher-order cognitive reasoning.** For instance, benchmarks like CA-MER [321], which assesses reasoning in emotionally conflicting scenarios, and CausalVQA [302], which systematically tests causal understanding, are designed to compel models to move beyond mere “perception” toward genuine “reasoning”. **Furthermore, task formats must involve open-ended knowledge integration and creation.** The evaluation emphasis should be on generating logically coherent explanations or executable code, rather than selecting a single correct answer. A forward-looking example is ChartMimic [292], which requires models to use visual, logical, and programming skills to reproduce a chart.

## 6 CONCLUSION

In this survey, we systematically track and summarize the evolution of vision–language interactive reasoning in Multimodal Large Language Models (MLLMs) under a unified From Perception to Cognition framework. To the best of our knowledge, this review offers one of the most comprehensive overviews of how fine-grained Perception underpins robust Cognition. Specifically, we first outline the developmental history of vision–language interaction and present essential background, including core concepts, representative architectures, and evaluation protocols. We then organize the field around two layers: **On the Perception side**, we synthesize advances in visual encoders, resolution handling, and task-aware vision–language alignment. **On the Cognition side**, we examine training paradigms for problem decomposition, preference optimization, and dynamic reasoning loops that re-inspect visual evidence. We also catalog benchmarks and applications across documents, charts, scientific reasoning, and healthcare, compile the relevant datasets, and offer fair, comparable analyses where feasible. Finally, we outline concrete future directions: building a truly unified, multi-granular visual encoder, developing latent-space reasoning that plans and searches over internal representations, advancing cross-image relational reasoning and integrating tool-augmented interaction. We also highlight the overarching challenge of closing the perception–cognition gap. This survey aims to serve both newcomers and experienced researchers as a structured, up-to-date reference on current progress and a guide for future work in vision–language interactive reasoning.

## REFERENCES

- [1] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He *et al.*, “Towards artificial general intelligence with hybrid tianjic chip architecture,” *Nature*, vol. 572, no. 7767, pp. 106–111, 2019. <sup>1</sup>
- [2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, “Sparks of artificial general intelligence: Early experiments with gpt-4,” *arXiv preprint arXiv:2303.12712*, 2023. <sup>1</sup>

- [3] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang *et al.*, "Towards artificial general intelligence via a multimodal foundation model," *Nature Communications*, vol. 13, no. 1, p. 3094, 2022. [1](#), [18](#)
- [4] J. N. Wood, "Artificial intelligence tackles the nature-nurture debate," *Nature Machine Intelligence*, vol. 6, no. 4, pp. 381–382, 2024. [1](#)
- [5] J. Y. Lee, S. Lee, A. Mishra, X. Yan, B. McMahan, B. Gaisford, C. Kobashigawa, M. Qu, C. Xie, and J. C. Kao, "Brain-computer interface control with artificial intelligence copilots," *Nature Machine Intelligence*, pp. 1–14, 2025. [1](#)
- [6] J. Xiong, W. Zhang, Y. Wang, J. Huang, Y. Shi, M. Xu, M. Li, Z. Fu, X. Kong, Y. Wang *et al.*, "Bridging chemistry and artificial intelligence by a reaction description language," *Nature Machine Intelligence*, pp. 1–12, 2025. [1](#)
- [7] S. Hu, J. Liu, Y. Wang, C. Fu, J. Zhu, H. Yu, and C. Yang, "Transparent artificial intelligence-enabled interpretable and interactive sleep apnea assessment across flexible monitoring scenarios," *Nature Communications*, vol. 16, no. 1, p. 7548, 2025. [1](#)
- [8] S. Tayebi Arasteh, T. Han, M. Lotfinia, C. Kuhl, J. N. Kather, D. Truhn, and S. Nebelung, "Large language models streamline automated machine learning for clinical studies," *Nature Communications*, vol. 15, no. 1, p. 1603, 2024. [1](#)
- [9] E. Kasneci, K. Seffler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023. [1](#)
- [10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023. [1](#)
- [11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. [1](#), [4](#), [5](#)
- [12] O. Simeoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025. [1](#), [5](#)
- [13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. [1](#)
- [14] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024. [1](#)
- [15] Q. Ye, Z. Yu, R. Shao, Y. Cui, X. Kang, X. Liu, P. Torr, and X. Cao, "Cat+: Investigating and enhancing audio-visual understanding in large language models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 10, pp. 8674–8690, 2025. [1](#), [2](#)
- [16] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Milligan, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022. [1](#), [8](#)
- [17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023. [1](#), [4](#), [8](#)
- [18] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy *et al.*, "Factuality challenges in the era of large language models and opportunities for fact-checking," *Nature Machine Intelligence*, vol. 6, no. 8, pp. 852–863, 2024. [1](#)
- [19] Y. Ma, L. Jiao, F. Liu, L. Li, W. Ma, S. Yang, X. Liu, and P. Chen, "Unveiling and mitigating generalized biases of dnns through the intrinsic dimensions of perceptual manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [1](#)
- [20] M. Griot, C. Hemptonne, J. Vanderdonckt, and D. Yuksel, "Large language models lack essential metacognition for reliable medical reasoning," *Nature communications*, vol. 16, no. 1, p. 642, 2025. [1](#)
- [21] J. Chen, Y. Ma, A. Zhang, W. Tang, W. Dai, and B. Liu, "Compositional attribute imbalance in vision datasets," *arXiv preprint arXiv:2506.14418*, 2025. [1](#)
- [22] Y. Xu, L. Hu, J. Zhao, Z. Qiu, K. Xu, Y. Ye, and H. Gu, "A survey on multilingual large language models: Corpora, alignment, and bias," *Frontiers of Computer Science*, vol. 19, no. 11, p. 1911362, 2025. [1](#)
- [23] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025. [1](#), [17](#), [19](#)
- [24] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024. [1](#)
- [25] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024. [1](#), [17](#)
- [26] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," *arXiv preprint arXiv:2509.04664*, 2025. [1](#)
- [27] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9568–9578. [2](#), [5](#), [6](#)
- [28] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024. [3](#)
- [29] Z.-Z. Li, D. Zhang, M.-L. Zhang, J. Zhang, Z. Liu, Y. Yao, H. Xu, J. Zheng, P.-J. Wang, X. Chen *et al.*, "From system 1 to system 2: A survey of reasoning large language models," *arXiv preprint arXiv:2502.17419*, 2025. [3](#)
- [30] Y. Wang, S. Wu, Y. Zhang, S. Yan, Z. Liu, J. Luo, and H. Fei, "Multimodal chain-of-thought reasoning: A comprehensive survey," *arXiv preprint arXiv:2503.12605*, 2025. [3](#)
- [31] Y. Li, Z. Liu, Z. Li, X. Zhang, Z. Xu, X. Chen, H. Shi, S. Jiang, X. Wang, J. Wang *et al.*, "Perception, reason, think, and plan: A survey on large multimodal reasoning models," *arXiv preprint arXiv:2505.04921*, 2025. [3](#)
- [32] F. Ke, J. Hsu, Z. Cai, Z. Ma, X. Zheng, X. Wu, S. Huang, W. Wang, P. D. Haghghi, G. Haffari *et al.*, "Explain before you answer: A survey on compositional visual reasoning," *arXiv preprint arXiv:2508.17298*, 2025. [3](#)
- [33] Z. Su, P. Xia, H. Guo, Z. Liu, Y. Ma, X. Qu, J. Liu, Y. Li, K. Zeng, Z. Yang *et al.*, "Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers," *arXiv preprint arXiv:2506.23918*, 2025. [3](#), [13](#)
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [4](#)
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [4](#)
- [36] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo, "Tap: Text-aware pre-training for text-vqa and text-caption," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8751–8761. [4](#)
- [37] H. Xia, R. Lan, H. Li, and S. Song, "St-vqa: shrinkage transformer with accurate alignment for visual question answering," *Applied Intelligence*, vol. 53, no. 18, pp. 20 967–20 978, 2023. [4](#)
- [38] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: a dataset for image captioning with reading comprehension," in *European conference on computer vision*. Springer, 2020, pp. 742–758. [4](#)
- [39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017. [4](#)
- [40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229. [4](#)
- [41] J. Kuang, Y. Shen, J. Xie, H. Luo, Z. Xu, R. Li, Y. Li, X. Cheng, X. Lin, and Y. Han, "Natural language understanding and inference with mllm in visual question answering: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–36, 2025. [4](#)
- [42] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 296–26 306. [4](#)
- [43] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023. [5](#), [6](#)
- [44] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF*

- international conference on computer vision*, 2023, pp. 11 975–11 986. 5, 6
- [45] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, “Demystifying clip data,” *arXiv preprint arXiv:2309.16671*, 2023. 5, 6
- [46] M. Oquab, T. Darcret, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023. 5, 6
- [47] W. Wang, Q. Sun, F. Zhang, Y. Tang, J. Liu, and X. Wang, “Diffusion feedback helps clip see better,” *arXiv preprint arXiv:2407.20171*, 2024. 5, 6
- [48] T. Zhang, Y. Li, Y.-c. Chou, J. Chen, A. Yuille, C. Wei, and J. Xiao, “Vision-language-vision auto-encoder: Scalable knowledge distillation from diffusion models,” *arXiv preprint arXiv:2507.07104*, 2025. 5, 6
- [49] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, “Prismatic vlms: Investigating the design space of visually-conditioned language models,” in *Forty-first International Conference on Machine Learning*, 2024. 5, 6
- [50] H. Zhang, H. You, P. Dufter, B. Zhang, C. Chen, H.-Y. Chen, T.-J. Fu, W. Y. Wang, S.-F. Chang, Z. Gan *et al.*, “Ferret-v2: An improved baseline for referring and grounding with large language models,” *arXiv preprint arXiv:2404.07973*, 2024. 5, 6
- [51] X. Fan, T. Ji, S. Li, S. Jin, S. Song, J. Wang, B. Hong, L. Chen, G. Zheng, M. Zhang *et al.*, “Poly-visual-expert vision-language models,” in *First Conference on Language Modeling*, 2024. 5, 6
- [52] O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, and F. Tombari, “Brave: Broadening the visual encoding of vision-language models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 113–132. 5, 6
- [53] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen *et al.*, “Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models,” *arXiv preprint arXiv:2311.07575*, 2023. 5, 6
- [54] A.-L. Wang, B. Shan, W. Shi, K.-Y. Lin, X. Fei, G. Tang, L. Liao, J. Tang, C. Huang, and W.-S. Zheng, “Pargo: Bridging vision-language with partial and global views,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7491–7499. 5, 6, 8
- [55] J. Lin, H. Chen, Y. Fan, Y. Fan, X. Jin, H. Su, J. Fu, and X. Shen, “Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4156–4166. 5, 6
- [56] L. Shen, G. Chen, R. Shao, W. Guan, and L. Nie, “Mome: Mixture of multimodal experts for generalist multimodal large language models,” *Advances in neural information processing systems*, vol. 37, pp. 42 048–42 070, 2024. 5, 6
- [57] Z. Zong, B. Ma, D. Shen, G. Song, H. Shao, D. Jiang, H. Li, and Y. Liu, “Mova: Adapting mixture of vision experts to multimodal context,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 103 305–103 333, 2024. 5, 6
- [58] W. Wang, X. Li, Z. Wang, Y. Pang, J. Zhang, P. Li, Q. Zhang, and L. Gao, “Diving into mitigating hallucinations from a vision perspective for large vision-language models,” *arXiv preprint arXiv:2509.13836*, 2025. 5, 6
- [59] Y. Wu, J. Du, K. Yan, S. Ding, and X. Li, “Tove: Efficient vision-language learning via knowledge transfer from vision experts,” *arXiv preprint arXiv:2504.00691*, 2025. 5, 6
- [60] Z. Li, Z. Li, and T. Zhou, “R2-t2: Re-routing in test-time for multimodal mixture-of-experts,” *arXiv preprint arXiv:2502.20395*, 2025. 5, 6
- [61] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov, “Am-radio: Agglomerative vision foundation model reduce all domains into one,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 12 490–12 500. 6
- [62] M. B. Sarıyıldız, P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis, “Unic: Universal classification models via multi-teacher distillation,” *arXiv preprint arXiv:2408.05088*, 2024. 6
- [63] J. Cao, Y. Zhang, T. Huang, M. Lu, Q. Zhang, R. An, N. Ma, and S. Zhang, “Move-kd: Knowledge distillation for vlms with mixture of visual encoders,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 846–19 856. 6
- [64] M. B. Sarıyıldız, P. Weinzaepfel, T. Lucas, P. de Jorge, D. Larlus, and Y. Kalantidis, “Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 084–30 094. 6
- [65] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” *arXiv preprint arXiv:2203.03605*, 2022. 5
- [66] G. Luo, Y. Zhou, Y. Zhang, X. Zheng, X. Sun, and R. Ji, “Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models,” *arXiv preprint arXiv:2403.03003*, 2024. 5
- [67] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, “Mini-gemini: Mining the potential of multi-modality vision language models,” *arXiv preprint arXiv:2403.18814*, 2024. 5
- [68] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, “Towards understanding the mixture-of-experts layer in deep learning,” *Advances in neural information processing systems*, vol. 35, pp. 23 049–23 062, 2022. 5, 8
- [69] J. Cha, W. Kang, J. Mun, and B. Roh, “Honeybee: Locality-enhanced projector for multimodal llm,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 817–13 827. 7, 8
- [70] X. Zhu, Y. Hu, F. Mo, M. Li, and J. Wu, “Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 81 225–81 256, 2024. 7, 8
- [71] Z. Xu, B. Qu, Y. Qi, S. Du, C. Xu, C. Yuan, and J. Guo, “Chartmoe: Mixture of diversely aligned expert connector for chart understanding,” *arXiv preprint arXiv:2409.03277*, 2024. 7, 8
- [72] H. Li, J. Chen, Z. Wei, S. Huang, T. Hui, J. Gao, X. Wei, and S. Liu, “Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8592–8603. 7, 8
- [73] J. Zhao, B. Sun, X. Chen, X. Wei, and Q. Hou, “Llava-octopus: Unlocking instruction-driven adaptive projector fusion for video understanding,” *arXiv preprint arXiv:2501.05067*, 2025. 7, 8
- [74] S. Lu, Y. Li, Y. Xia, Y. Hu, S. Zhao, Y. Ma, Z. Wei, Y. Li, L. Duan, J. Zhao *et al.*, “Ovis2. 5 technical report,” *arXiv preprint arXiv:2508.11737*, 2025. 7, 8
- [75] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. M. Eisenschlos, “Matcha: Enhancing visual language pretraining with math reasoning and chart derendering,” *arXiv preprint arXiv:2212.09662*, 2022. 7, 8
- [76] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023. 7, 8, 18
- [77] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai *et al.*, “Q-instruct: Improving low-level visual abilities for multi-modality foundation models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 490–25 500. 7, 8
- [78] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty, “Chartinstruct: Instruction tuning for chart comprehension and reasoning,” *arXiv preprint arXiv:2403.09028*, 2024. 7, 8, 10
- [79] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, “Visual prompt tuning,” in *European conference on computer vision*. Springer, 2022, pp. 709–727. 7, 8
- [80] A. Zhang, H. Fei, Y. Yao, W. Ji, L. Li, Z. Liu, and T.-S. Chua, “Vpgtrans: Transfer visual prompt generator across llms,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 20 299–20 319, 2023. 7, 8
- [81] Y. Zhang, Y. Dong, S. Zhang, T. Min, H. Su, and J. Zhu, “Exploring the transferability of visual prompting for multimodal large language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 562–26 572. 7, 8
- [82] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, “Shikra: Unleashing multimodal llm’s referential dialogue magic,” *arXiv preprint arXiv:2306.15195*, 2023. 7, 8
- [83] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, Leizhang, C. Li *et al.*, “Llava-grounding: Grounded visual chat with large multimodal models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 19–35. 7, 8
- [84] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, Q. Ye, and F. Wei, “Grounding multimodal large language models to the world,” in *The Twelfth International Conference on Learning Representations*, 2024. 7, 8
- [85] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan, “Glamm: Pixel grounding large multimodal model,” in *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 009–13 018. 7, 8
- [86] M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and Y. J. Lee, “Vip-llava: Making large multimodal models understand arbitrary visual prompts,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 914–12 923. 7, 8
- [87] W. Lin *et al.*, “Draw-and-understand: Leveraging visual prompts to enable multimodal large language models,” *arXiv preprint arXiv:2403.20271*, 2024. 7, 8
- [88] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, “Lisa: Reasoning segmentation via large language model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589. 7, 9
- [89] Z. Xia, D. Han, Y. Han, X. Pan, S. Song, and G. Huang, “Gsva: Generalized segmentation via multimodal large language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3858–3869. 7, 9
- [90] J. Wu, M. Zhong, S. Xing, Z. Lai, Z. Liu, Z. Chen, W. Wang, X. Zhu, L. Lu, T. Lu, P. Luo, Y. Qiao, and J. Dai, “Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks,” *arXiv preprint arXiv:2406.08394*, 2024. 7, 9
- [91] H. Fei, S. Wu, H. Zhang, T.-S. Chua, and S. Yan, “Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing,” *Advances in neural information processing systems*, vol. 37, pp. 57 207–57 239, 2024. 7, 9
- [92] D. Jang *et al.*, “A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation,” *arXiv preprint arXiv:2503.13881*, 2025. 7, 9
- [93] L. Zhu, T. Chen, Q. Xu, X. Liu, D. Ji, H. Wu, D. W. Soh, and J. Liu, “Popen: Preference-based optimization and ensemble for lilm-based reasoning segmentation,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 231–30 240. 7, 9
- [94] P. Wu and S. Xie, “V\*: Guided visual search as a core mechanism in multimodal llms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 084–13 094. 7, 9, 22
- [95] G. Li, J. Xu, Y. Zhao, and Y. Peng, “Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9098–9108. 7, 9, 22
- [96] G. Sun, M. Jin, Z. Wang, C.-L. Wang, S. Ma, Q. Wang, T. Geng, Y. N. Wu, Y. Zhang, and D. Liu, “Visual agents as fast and slow thinkers,” *arXiv preprint arXiv:2408.08862*, 2024. 7, 9
- [97] H. Taud and J.-F. Mas, “Multilayer perceptron (mlp),” in *Geometric approaches for modeling land change scenarios*. Springer, 2017, pp. 451–455. 8
- [98] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742. 8
- [99] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023. 8, 18
- [100] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye, “Ovis: Structural embedding alignment for multimodal large language model,” *arXiv preprint arXiv:2405.20797*, 2024. 8
- [101] W. Shi, Z. Hu, Y. Bin, J. Liu, Y. Yang, S.-K. Ng, L. Bing, and R. K.-W. Lee, “Math-llava: Bootstrapping mathematical reasoning for multimodal large language models,” *arXiv preprint arXiv:2406.17294*, 2024. 8
- [102] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty, “Chartgemma: Visual instruction-tuning for chart reasoning in the wild,” *arXiv preprint arXiv:2407.04172*, 2024. 8, 9, 10
- [103] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, “Pixelmm: Pixel reasoning with large multimodal model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 374–26 383. 8
- [104] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, “Let’s verify step by step,” in *The Twelfth International Conference on Learning Representations*, 2023. 9, 10
- [105] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023. 9, 10, 14
- [106] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, “Dolphins: Multimodal language model for driving,” in *European Conference on Computer Vision*. Springer, 2024, pp. 403–420. 9, 10
- [107] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, “Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 8612–8642, 2024. 9, 10
- [108] G. Xu, P. Jin, Z. Wu, H. Li, Y. Song, L. Sun, and L. Yuan, “Llava-cot: Let vision language models reason step-by-step,” *arXiv preprint arXiv:2411.10440*, 2024. 9, 10
- [109] O. Thawakar, D. Dissanayake, K. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, M. Zumri, J. Lahoud, R. M. Anwer *et al.*, “Llamav-01: Rethinking step-by-step visual reasoning in llms,” *arXiv preprint arXiv:2501.06186*, 2025. 9, 10
- [110] Y. Sheng, L. Li, and D. D. Zeng, “Learning theorem rationale for improving the mathematical reasoning capability of large language models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 15 151–15 159. 9, 10
- [111] K. Zhao, B. Zhu, Q. Sun, and H. Zhang, “Unsupervised visual chain-of-thought reasoning via preference optimization,” *arXiv preprint arXiv:2504.18397*, 2025. 9, 10
- [112] J. Wang, Z. Kang, H. Wang, H. Jiang, J. Li, B. Wu, Y. Wang, J. Ran, X. Liang, C. Feng, and J. Xiao, “Vgr: Visual grounded reasoning,” *arXiv preprint arXiv:2506.11991*, 2025. 9, 10
- [113] H. Shen, P. Liu, J. Li, C. Fang, Y. Ma, J. Liao, Q. Shen, Z. Zhang, K. Zhao, Q. Zhang, R. Xu, and T. Zhao, “Vlm-r1: A stable and generalizable r1-style large vision-language model,” *arXiv preprint arXiv:2504.07615*, 2025. 9, 10
- [114] Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang, “Visual-rft: Visual reinforcement fine-tuning,” *arXiv preprint arXiv:2503.01785*, 2025. 9, 10
- [115] L. Zhao, E. Yu, Z. Ge, J. Yang, H. Wei, H. Zhou, J. Sun, Y. Peng, R. Dong, C. Han *et al.*, “Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning,” *arXiv preprint arXiv:2307.09474*, 2023. 8
- [116] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu, “Osprey: Pixel understanding with visual instruction tuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 202–28 211. 8
- [117] A. Shtedritski, C. Rupprecht, and A. Vedaldi, “What does clip know about a red circle? visual prompt engineering for vlms,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 987–11 997. 8
- [118] X. Li, H. Yuan, W. Li, H. Ding, S. Wu, W. Zhang, Y. Li, K. Chen, and C. C. Loy, “Omg-seg: Is one model good enough for all segmentation?” *arXiv preprint arXiv:2401.10229*, 2024. 8
- [119] J. Wang and L. Ke, “Llm-seg: Bridging image segmentation and large language model reasoning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1765–1774. 9
- [120] T. Zhang, X. Li, H. Fei, H. Yuan, S. Wu, S. Ji, C. C. Loy, and S. Yan, “Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding,” *Advances in neural information processing systems*, vol. 37, pp. 71 737–71 767, 2024. 9
- [121] J. Zheng, J. Li, S. Cheng, Y. Zheng, J. Li, J. Liu, Y. Liu, J. Liu, and X. Zhan, “Instruction-guided visual masking,” *Advances in neural information processing systems*, vol. 37, pp. 126 004–126 031, 2024. 9
- [122] L. Zhu, T. Chen, D. Ji, J. Ye, and J. Liu, “Llfafs: When large language models meet few-shot segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3065–3075. 9
- [123] Y. Zhang, S. Qian, B. Peng, S. Liu, and J. Jia, “Prompt highlighter: Interactive control for multi-modal llms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 215–13 224. 9
- [124] R. Pi, L. Yao, J. Gao, J. Zhang, and T. Zhang, “Perceptiongpt: Effectively fusing visual perception into llm,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 27 124–27 133. 9
- [125] Y. Shen, C. Fu, P. Chen, M. Zhang, K. Li, X. Sun, Y. Wu, S. Lin, and R. Ji, “Aligning and prompting everything all at once for universal visual perception,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 13 193–13 203. 9
- [126] Z. Lin, Y. Gao, X. Zhao, Y. Yang, and J. Sang, “Mind with eyes:

- from language reasoning to multimodal reasoning," *arXiv preprint arXiv:2503.18071*, 2025. 9
- [127] L. Guo, A. C. Rivera, P. Tang, H. Ren, and Z. Song, "Hierarchical contextual grounding lvm: Enhancing fine-grained visual-language understanding with robust grounding," *arXiv preprint arXiv:2508.16974*, 2025. 9
- [128] C. B. Browne, E. P. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012. 9, 22
- [129] Z. Ma, Y. Zhou, Z. Wang, B. Ding, and J. Luo, "Crepe: Can vision-language foundation models reason compositionally?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10910–10920. 9
- [130] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang, "Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023. 9
- [131] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022. 9
- [132] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022. 9, 10, 11
- [133] Y. Wang, S. Wu, Y. Zhang, W. Y. Wang, Z. Liu, J. Luo, and H. Fei, "Multimodal chain-of-thought reasoning: A comprehensive survey," *arXiv preprint arXiv:2503.12605*, 2025. 9
- [134] R. Zhao, Q. Yuan, J. Li, H. Hu, Y. Li, C. Zheng, and F. Gao, "Sce2drivex: A generalized mllm framework for scene-to-drive learning," *arXiv preprint arXiv:2502.14917*, 2025. 10
- [135] Z. Shao *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024. 10
- [136] H. Tan, Y. Ji, X. Hao, M. Lin, P. Wang, Z. Wang, and S. Zhang, "Reason-rft: Reinforcement fine-tuning for visual reasoning," *arXiv preprint arXiv:2503.20752*, 2025. 10
- [137] Y. Liu, B. Peng, Z. Zhong, Z. Yue, F. Lu, B. Yu, and J. Jia, "Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement," *arXiv preprint arXiv:2503.06520*, 2025. 10
- [138] Y. Yang, X. He, H. Pan, X. Jiang, Y. Deng, X. Yang, H. Lu, D. Yin, F. Rao, M. Zhu *et al.*, "R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization," *arXiv preprint arXiv:2503.10615*, 2025. 10
- [139] Z. Wang, K. Wang, Q. Wang, P. Zhang, L. Li, Z. Yang, X. Jin, K. Yu, M. N. Nguyen, L. Liu *et al.*, "Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning," *arXiv preprint arXiv:2504.20073*, 2025. 10
- [140] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *arXiv preprint arXiv:2305.18290*, 2023. 10, 12
- [141] Y. Xie, G. Li, X. Xu, and M.-Y. Kan, "V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization," *arXiv preprint arXiv:2411.02712*, 2024. 10
- [142] T. Bai, Z. Hu, F. Sun, J. Qiu, Y. Jiang, G. He, B. Zeng, C. He, B. Yuan, and W. Zhang, "Multi-step visual reasoning with visual tokens scaling and verification," *arXiv preprint arXiv:2506.07235*, 2025. 10
- [143] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Advances in Neural Information Processing Systems*, 2022. 11
- [144] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," in *European conference on computer vision*. Springer, 2022, pp. 146–162. 11
- [145] J. Qi, M. Ding, W. Wang, Y. Bai, Q. Lv, W. Hong, B. Xu *et al.*, "Cogcom: A visual language model with chain-of-manipulations," *arXiv preprint arXiv:2402.04236*, 2024. 11, 14
- [146] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM journal on computing*, vol. 1, no. 2, pp. 146–160, 1972. 11
- [147] X. Chen, R. Zhang, D. Jiang, A. Zhou, S. Yan, W. Lin, and H. Li, "Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning," *arXiv preprint arXiv:2506.05331*, 2025. 11, 14
- [148] H. Yao *et al.*, "Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search," *arXiv preprint arXiv:2412.18319*, 2024. 11
- [149] Q. Wu, X. Yang, Y. Zhou, C. Fang, B. Song, X. Sun, and R. Ji, "Grounded chain-of-thought for multimodal large language models," *arXiv preprint arXiv:2503.12799*, 2025. 11
- [150] Z. Chen, R. Zhao, C. Luo, M. Sun, X. Yu, Y. Kang, and R. Huang, "Sifthinker: Spatially-aware image focus for visual reasoning," *arXiv preprint arXiv:2508.06259*, 2025. 11, 14
- [151] A. Su, H. Wang, W. Ren, F. Lin, and W. Chen, "Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning," *arXiv preprint arXiv:2505.15966*, 2025. 11, 14, 22
- [152] Z. Ma, Z. Liu, J. Zhang, J. Tan, M. Shu, J. C. Niebles, S. Heinecke, H. Wang, C. Xiong, R. Krishna, and S. Savarese, "Latte: Learning to think with vision specialists," *arXiv preprint arXiv:2412.05479*, 2024. 11
- [153] Z. Ma, J. Zhang, Z. Liu, J. Zhang, J. Tan, M. Shu, J. C. Niebles, S. Heinecke, H. Wang, C. Xiong, R. Krishna, and S. Savarese, "Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action," *arXiv preprint arXiv:2412.05479*, 2024. 11, 15
- [154] E. Zelikman, Y. Wu, N. D. Goodman, and O. Holzman, "Star: Bootstrapping reasoning with reasoning," *arXiv preprint arXiv:2203.14465*, 2022. 12
- [155] C. Tan, J. Wei, Z. Gao, L. Sun, S. Li, X. Yang, and S. Z. Li, "Boosting the power of small multimodal reasoning models to match larger models with self-consistency training," *arXiv preprint arXiv:2311.14109*, 2023. 12
- [156] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022. 12, 13
- [157] J. Xia, B. Tong, Y. Zang, R. Shao, and K. Zhou, "Bootstrapping grounded chain-of-thought in multimodal llms for data-efficient model adaptation," *arXiv preprint arXiv:2507.02859*, 2025. 12
- [158] T. Yu, H. Li, R. Zhou, J. Chen, C. Zhang, Y. Jiang, and D. Tao, "Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness," *arXiv preprint arXiv:2405.17220*, 2024. 12
- [159] R. Zhang, B. Zhang, Y. Li, H. Zhang, Z. Sun, Z. Gan, Y. Yang, R. Pang, and Y. Yang, "Improve vision language model chain-of-thought reasoning," *arXiv preprint arXiv:2410.16198*, 2024. 12
- [160] K. Yang, D. Klein, A. Celikyilmaz, N. Peng, and Y. Tian, "Rlcd: Reinforcement learning from contrastive distillation for language model alignment," *arXiv preprint arXiv:2307.12950*, 2023. 12
- [161] J. He, H. Lin, Q. Wang, Y. Fung, and H. Ji, "Self-correction is more than refinement: A learning framework for visual and language reasoning tasks," *arXiv preprint arXiv:2410.04055*, 2024. 12
- [162] R. Pi, T. Han, W. Xiong, J. Zhang, R. Liu, R. Pan, and T. Zhang, "Strengthening multimodal large language model with bootstrapped preference optimization," in *European Conference on Computer Vision*. Springer, 2024, pp. 382–398. 12
- [163] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023. 13
- [164] A. Bundy and L. Wallen, "Breadth-first search," in *Catalogue of artificial intelligence tools*. Springer, 1984, pp. 13–13. 13
- [165] L. Gui, C. Gârbacea, and V. Veitch, "Bonbon alignment for large language models and the sweetness of best-of-n sampling," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2851–2885, 2024. 13
- [166] Y. Xie, K. Kawaguchi, Y. Zhao, J. X. Zhao, M.-Y. Kan, J. He, and M. Xie, "Self-evaluation guided beam search for reasoning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 41 618–41 650, 2023. 13
- [167] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019. 13
- [168] G. Chen, M. Liao, C. Li, and K. Fan, "Alphamat almost zero: process supervision without process," *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 689–27 724, 2024. 13
- [169] D. Zhang, S. Zhoubian, Z. Hu, Y. Yue, Y. Dong, and J. Tang, "Rest-mcts\*: Llm self-training via process reward guided tree search," *Advances in Neural Information Processing Systems*, vol. 37, pp. 64 735–64 772, 2024. 13

- [170] G. Chen, M. Liao, C. Li, and K. Fan, "Step-level value preference optimization for mathematical reasoning," *arXiv preprint arXiv:2406.10858*, 2024. [13](#)
- [171] Y. Wang, S. Wang, Q. Cheng, Z. Fei, L. Ding, Q. Guo, D. Tao, and X. Qiu, "Visuothink: Empowering lvm reasoning with multimodal tree search," *arXiv preprint arXiv:2504.09130*, 2025. [13](#)
- [172] D. Acuna, X. Lu, J. Jung, H. Kim, A. Kar, S. Fidler, and Y. Choi, "Socratic-mcts: Test-time visual reasoning by asking the right questions," *arXiv preprint arXiv:2506.08927*, 2025. [13](#)
- [173] C. Zhang, J. Peng, Z. Wang, Y. Lai, H. Sun, H. Chang, F. Ma, and W. Yu, "Vrest: Enhancing reasoning in large vision-language models through tree search and self-reward mechanism," *arXiv preprint arXiv:2506.08691*, 2025. [13](#)
- [174] J. Wu, M. Feng, S. Zhang, R. Jin, F. Che, Z. Wen, and J. Tao, "Boosting multimodal reasoning with mcts-automated structured thinking," *arXiv e-prints*, pp. arXiv-2502, 2025. [13](#)
- [175] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 418–13 427. [13](#)
- [176] J. Li, J. Zhang, Z. Jie, L. Ma, and G. Li, "Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding," *arXiv preprint arXiv:2501.01926*, 2025. [13](#)
- [177] Q. Hu, A. Wang, J. Song, D. Qiu, Q. Liu, and J. Su, "Boosting visual knowledge-intensive training for lvms through causality-driven visual object completion," *arXiv preprint arXiv:2508.04453*, 2025. [14](#)
- [178] J. Gao, Y. Li, Z. Cao, and W. Li, "Interleaved-modal chain-of-thought," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 520–19 529. [14](#)
- [179] S. Yang, Y. Niu, Y. Liu, Y. Ye, B. Lin, and L. Yuan, "Look-back: Implicit visual re-focusing in mllm reasoning," *arXiv preprint arXiv:2507.03019*, 2025. [14](#)
- [180] Z. Zheng, M. Yang, J. Hong, C. Zhao, G. Xu, L. Yang, C. Shen, and X. Yu, "Deepeyes: Incentivizing" thinking with images" via reinforcement learning," *arXiv preprint arXiv:2505.14362*, 2025. [14, 22](#)
- [181] G. Zhang, T. Zhong, Y. Xia, Z. Yu, H. Li, W. He, F. Shu, M. Liu, D. She, Y. Wang *et al.*, "Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation," *arXiv preprint arXiv:2503.05255*, 2025. [14, 22](#)
- [182] M. Ni, Z. Yang, L. Li, C.-C. Lin, K. Lin, W. Zuo, and L. Wang, "Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning," *arXiv preprint arXiv:2505.19702*, 2025. [14](#)
- [183] J. Chung, J. Kim, S. Kim, J. Lee, M. S. Kim, and Y. Yu, "Don't look only once: Towards multimodal interactive reasoning with selective visual revisit," *arXiv preprint arXiv:2505.18842*, 2025. [14](#)
- [184] B. Luan, H. Feng, H. Chen, Y. Wang, W. Zhou, and H. Li, "Textcot: Zoom in for enhanced multimodal text-rich image understanding," *arXiv preprint arXiv:2404.09797*, 2024. [14](#)
- [185] Y. Hu, W. Shi, X. Fu, D. Roth, M. Ostendorf, L. Zettlemoyer, N. A. Smith, and R. Krishna, "Visual sketchpad: Sketching as a visual chain of thought for multimodal language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 348–139 379, 2024. [14, 15](#)
- [186] Z. Liu, Y. Dong, Y. Rao, J. Zhou, and J. Lu, "Chain-of-spot: Interactive reasoning improves large vision-language models," *arXiv preprint arXiv:2403.12966*, 2024. [14](#)
- [187] Q.-H. Le, L. H. Dang, N. H. Le, T. Tran, and T. M. Le, "Progressive multi-granular alignments for grounded reasoning in large vision-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4473–4481. [14](#)
- [188] Z. Li, R. Luo, J. Zhang, M. Qiu, X. Huang, and Z. Wei, "Vocot: Unleashing visually grounded multi-step reasoning in large multimodal models," *arXiv preprint arXiv:2405.16919*, 2024. [14](#)
- [189] M. Zhu, H. Zhong, C. Zhao, Z. Du, Z. Huang, M. Liu, H. Chen, C. Zou, J. Chen, M. Yang *et al.*, "Active-o3: Empowering multimodal large language models with active perception via grpo," *arXiv preprint arXiv:2505.21457*, 2025. [14](#)
- [190] C. Jiang, Y. Heng, W. Ye, H. Yang, H. Xu, M. Yan, J. Zhang, F. Huang, and S. Zhang, "VLM-R<sup>3</sup>: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought," *arXiv preprint arXiv:2505.16192*, 2025. [14](#)
- [191] S. Wang, G. Fang, L. Kong, X. Li, J. Xu, S. Yang, Q. Li, J. Zhu, and X. Wang, "Pixelthink: Towards efficient chain-of-pixel reasoning," *arXiv preprint arXiv:2505.23727*, 2025. [14](#)
- [192] M. Bigverdi, Z. Luo, C.-Y. Hsieh, E. Shen, D. Chen, L. G. Shapiro, and R. Krishna, "Perception tokens enhance visual reasoning in multimodal language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3836–3845. [14](#)
- [193] Y. Man, D.-A. Huang, G. Liu, S. Sheng, S. Liu, L.-Y. Gui, J. Kautz, Y.-X. Wang, and Z. Yu, "Argus: Vision-centric reasoning with grounded chain-of-thought," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 268–14 280. [14](#)
- [194] Z. Yang, Y. Zhang, X. Wang, L. Yang, J. Wang, L. Zhang, Y. Zhang, Z. Liu, J. Gao, L. Wang *et al.*, "Mm-react: Prompting chatgpt for multimodal reasoning and action," *arXiv preprint arXiv:2303.11381*, 2023. [15](#)
- [195] D. Suris, S. Menon, and C. Vondrick, "Vipergpt: Visual inference via python execution for reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 888–11 898. [14, 15](#)
- [196] Z. Gao *et al.*, "Clova: A closed-loop visual assistant with tool usage and update," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [15](#)
- [197] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu *et al.*, "Llava-plus: Learning to use tools for creating multimodal agents," in *European conference on computer vision*. Springer, 2024, pp. 126–142. [15](#)
- [198] Z. Su *et al.*, "Openthinkimg: Learning to think with images via visual tool reinforcement learning," *arXiv preprint arXiv:2505.08617*, 2025. [15, 22](#)
- [199] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 14 953–14 962. [14](#)
- [200] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38 154–38 180, 2023. [14](#)
- [201] Z. Liu, Y. He, W. Wang, W. Wang, Y. Wang, S. Chen, Q. Zhang, Z. Lai, Y. Yang, Q. Li *et al.*, "Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language," *arXiv preprint arXiv:2305.05662*, 2023. [14](#)
- [202] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022. [15](#)
- [203] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023. [15](#)
- [204] Z. Khan, V. K. BG, S. Schulter, Y. Fu, and M. Chandraker, "Self-training large language models for improved visual program synthesis with visual reinforcement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 344–14 353. [15](#)
- [205] Z. Ma, Z. Liu, J. Zhang, J. Tan, M. Shu, J. C. Niebles, S. Heinecke, H. Wang, C. Xiong, R. Krishna, and S. Savarese, "Latte: Learning to reason with vision specialists," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. [15](#)
- [206] F. Ke, X. Leng, Z. Cai, Z. Khan, W. Wang, P. D. Haghighi, H. Rezatofighi, M. Chandraker *et al.*, "Dwim: Towards tool-aware visual reasoning via discrepancy-aware workflow generation & instruct-masking tuning," *arXiv preprint arXiv:2503.19263*, 2025. [15](#)
- [207] X. Fu, M. Liu, Z. Yang, J. Corring, Y. Lu, J. Yang, D. Roth, D. Florencio, and C. Zhang, "Refocus: Visual editing as a chain of thought for structured image understanding," *arXiv preprint arXiv:2501.05452*, 2025. [15](#)
- [208] Y. Hu, O. Stretcu, C.-T. Lu, K. Viswanathan, K. Hata, E. Luo, R. Krishna, and A. Fuxman, "Visual programs distillation: Distilling tools and programmatic reasoning into vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9590–9601. [15](#)
- [209] C. Wang, W. Luo, S. Dong, X. Xuan, Z. Li, L. Ma, and S. Gao, "Mllm-tool: A multimodal large language model for tool agent learning," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 6678–6687. [15](#)

- [210] M. Wu, J. Yang, J. Jiang, M. Li, K. Yan, H. Yu, M. Zhang, C. Zhai, and K. Nahrstedt, "Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use," *arXiv preprint arXiv:2505.19255*, 2025. [15](#)
- [211] Z. Liu *et al.*, "Visual agentic reinforcement fine-tuning," *arXiv preprint arXiv:2505.14246*, 2025. [15](#)
- [212] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433. [15](#)
- [213] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913. [15](#)
- [214] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709. [15](#)
- [215] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731. [15](#)
- [216] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Okvqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204. [15](#)
- [217] J. Gao, Q. Wu, A. Blair, and M. Pagnucco, "Lora: A logical reasoning augmented dataset for visual question answering," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 579–30 591, 2023. [15](#)
- [218] L. M. Schulze Buschoff, E. Akata, M. Bethge, and E. Schulz, "Visual cognition in multimodal large language models," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 96–106, 2025. [16](#), [17](#)
- [219] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, "Raven: A dataset for relational and analogical visual reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5317–5327. [16](#)
- [220] D. Ghosal, V. T. Y. Han, C. Y. Ken, and S. Poria, "Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning," *arXiv preprint arXiv:2403.03864*, 2024. [16](#)
- [221] X. Liu, D. Yin, Y. Feng, and D. Zhao, "Things not written in text: Exploring spatial commonsense from visual signals," *arXiv preprint arXiv:2203.08075*, 2022. [16](#)
- [222] E. Liu, C. Cui, K. Zheng, and G. Neubig, "Testing the ability of language models to interpret figurative language," *arXiv preprint arXiv:2204.12632*, 2022. [16](#)
- [223] N. Alampara, M. Schilling-Wilhelmi, M. Ríos-García, I. Mandal, P. Khetarpal, H. S. Grover, N. A. Krishnan, and K. M. Jablonka, "Probing the limitations of multimodal language models for chemistry and materials research," *Nature computational science*, pp. 1–10, 2025. [16](#)
- [224] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022. [16](#)
- [225] M. Dreyer, J. Berend, T. Labarta, J. Vielhaben, T. Wiegand, S. Lapuschkin, and W. Samek, "Mechanistic understanding and validation of large ai models with semanticlens," *Nature Machine Intelligence*, pp. 1–14, 2025. [16](#), [17](#)
- [226] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," *arXiv preprint arXiv:2310.02255*, 2023. [16](#)
- [227] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao *et al.*, "Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?" in *European Conference on Computer Vision*. Springer, 2024, pp. 169–186. [16](#)
- [228] K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li, "Measuring multimodal mathematical reasoning with math-vision dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 095–95 169, 2024. [16](#)
- [229] P. Zhou, F. Zhang, X. Peng, Z. Xu, J. Ai, Y. Qiu, C. Li, Z. Li, M. Li, Y. Feng *et al.*, "Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models," *arXiv preprint arXiv:2504.05782*, 2025. [16](#)
- [230] C. Ma, J. Ding, J. Zhang, Z. Ma, H. Qing, B. Gao, L. Chen, M. Song *et al.*, "Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas," *arXiv preprint arXiv:2504.06637*, 2025. [16](#)
- [231] J. Feng, Z. Wang, Z. Zhang, Y. Guo, Z. Zhou, X. Chen, Z. Li, and D. Yin, "Mathreal: We keep it real! a real scene benchmark for evaluating math reasoning in multimodal large language models," *arXiv preprint arXiv:2508.06009*, 2025. [16](#)
- [232] Z. Zhou, S. Liu, M. Ning, W. Liu, J. Wang, D. F. Wong, X. Huang, Q. Wang, and K. Huang, "Is your model really a good math reasoner? evaluating mathematical reasoning with checklist," *arXiv preprint arXiv:2407.08733*, 2024. [16](#)
- [233] X. Xu, J. Zhang, T. Chen, Z. Chao, J. Hu, and C. Yang, "Ugmath-bench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models," *arXiv preprint arXiv:2501.13766*, 2025. [16](#)
- [234] Y. Hao, J. Gu, H. W. Wang, L. Li, Z. Yang, L. Wang, and Y. Cheng, "Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark," *arXiv preprint arXiv:2501.05444*, 2025. [16](#)
- [235] P. Wang, Z.-Z. Li, F. Yin, D. Ran, and C.-L. Liu, "Mv-math: Evaluating multimodal math reasoning in multi-visual contexts," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 541–19 551. [16](#)
- [236] Z. Hu, J. Liu, Z. Liu, Y. Liu, Z. Xie, and Y. Song, "Rmath: A logic reasoning-focused datasets toward mathematical multistep reasoning tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 24 104–24 112. [16](#)
- [237] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567. [16](#)
- [238] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021. [16](#)
- [239] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset, 2021," URL <https://arxiv.org/abs/2103.03874>, vol. 2, 2024. [16](#)
- [240] M.-H. Guo, J. Xu, Y. Zhang, J. Song, H. Peng, Y.-X. Deng, X. Dong, K. Nakayama, Z. Geng, C. Wang *et al.*, "R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation," *arXiv preprint arXiv:2505.02018*, 2025. [16](#)
- [241] J. Zhang, C. Petru, K. Nikolić, and F. Tramèr, "Realmath: A continuous benchmark for evaluating language models on research-level mathematics," *arXiv preprint arXiv:2505.12575*, 2025. [16](#)
- [242] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025. [17](#)
- [243] Google, "Introducing gemini 2.0: Our new ai model for the agentic era," <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>, Dec 2024, accessed: 2025-09-24. [17](#)
- [244] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024. [17](#)
- [245] OpenAI, "Gpt-5 system card," OpenAI, Tech. Rep., 2025, accessed: 2025-09-24. [Online]. Available: <https://cdn.openai.com/gpt-5-system-card.pdf> [17](#)
- [246] —, "Gpt-4 technical report," OpenAI, Tech. Rep., March 2023, gPT-4 Turbo was announced later as an updated version. Accessed: 2025-09-24. [Online]. Available: <https://cdn.openai.com/papers/gpt-4.pdf> [17](#)
- [247] —, "Gpt-4v system card," OpenAI, Tech. Rep., 2024, accessed: 2025-09-24. [Online]. Available: [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf) [17](#)
- [248] xAI, "Grok-3 beta release," <https://x.ai/news/grok-3>, 2025, accessed: 2025-09-24. [17](#)
- [249] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024. [17](#)

- [250] Anthropic, "Introducing claude 3.5 sonnet," <https://www.anthropic.com/news/clause-3-5-sonnet>, June 2024, accessed: 2025-09-24. 17
- [251] Qwen Team, "Qwen2.5-max: Exploring the intelligence of large-scale moe model," <https://qwenlm.github.io/blog/qwen2.5-max/>, January 2025, accessed: 2025-09-24. 17
- [252] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao *et al.*, "Kimi k1. 5: Scaling reinforcement learning with llms," *arXiv preprint arXiv:2501.12599*, 2025. 17
- [253] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025. 17
- [254] Qwen Team, "Qvq: To see the world with wisdom," <https://qwenlm.github.io/blog/qvq-72b-preview/>, December 2024, accessed: 2025-09-24. 17
- [255] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024. 17
- [256] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv-2407, 2024. 17
- [257] W. Shao *et al.*, "Ovis2.5 technical report," *arXiv preprint arXiv:2508.11737*, 2025. [Online]. Available: <https://arxiv.org/abs/2508.11737> 17
- [258] L. Du, F. Meng, Z. Liu, Z. Zhou, P. Luo, Q. Zhang, and W. Shao, "Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision," *arXiv preprint arXiv:2505.13427*, 2025. 16
- [259] G. Zhou, P. Qiu, C. Chen, J. Wang, Z. Yang, J. Xu, and M. Qiu, "Reinforced mllm: A survey on rl-based reasoning in multimodal large language models," *arXiv preprint arXiv:2504.21277*, 2025. 16
- [260] B. P. de Almeida, G. Richard, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, C. Rajesh, M. Lopez, A. Laterre *et al.*, "A multimodal conversational agent for dna, rna and protein tasks," *Nature Machine Intelligence*, pp. 1–14, 2025. 17
- [261] Q. Wu, P. Wang, X. Wang, X. He, and W. Zhu, "Medical vqa," in *Visual Question Answering: From Theory to Application*. Springer, 2022, pp. 165–176. 17
- [262] J. Wu, Y. Kim, and H. Wu, "Hallucination benchmark in medical visual question answering," *arXiv preprint arXiv:2401.05827*, 2024. 17
- [263] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "Pathvqa: 30000+ questions for medical visual question answering," *arXiv preprint arXiv:2003.10286*, 2020. 17, 22
- [264] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, and M. A. Riegler, "Kvasir-vqa: A text-image pair git tract dataset," in *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, 2024, pp. 3–12. 17
- [265] L. Bai, M. Islam, L. Seenivasan, and H. Ren, "Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery," *arXiv preprint arXiv:2305.11692*, 2023. 17
- [266] J. Burgess, J. J. Nirschl, L. Bravo-Sánchez, A. Lozano, S. R. Gupte, J. G. Galaz-Montoya, Y. Zhang, Y. Su, D. Bhowmik, Z. Coman *et al.*, "Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 552–19 564. 17
- [267] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Development of a large-scale medical visual question-answering dataset," *Communications Medicine*, vol. 4, no. 1, p. 277, 2024. 17
- [268] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, 2021, pp. 1650–1654. 17
- [269] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Pmc-vqa: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023. 17, 18
- [270] K. Zuo and Y. Jiang, "Medhallbench: A new benchmark for assessing hallucination in medical large language models," *arXiv preprint arXiv:2412.18947*, 2024. 17
- [271] J. Chen, D. Yang, T. Wu, Y. Jiang, X. Hou, M. Li, S. Wang, D. Xiao, K. Li, and L. Zhang, "Detecting and evaluating medical hallucinations in large vision language models," *arXiv preprint arXiv:2406.10185*, 2024. 17
- [272] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. A10a2300138, 2024. 18
- [273] A. Ing, A. Andrades, M. R. Cosenza, and J. O. Korbel, "Integrating multimodal cancer data using deep latent variable path modelling," *Nature Machine Intelligence*, pp. 1–23, 2025. 18
- [274] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024. 18
- [275] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. 18
- [276] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, S. XiXuan *et al.*, "Cogvilm: Visual expert for pretrained language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 475–121 499, 2024. 18
- [277] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367. 18
- [278] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023. 18
- [279] Y. Chen, D. Xu, Y. Huang, S. Zhan, H. Wang, D. Chen, X. Wang, M. Qiu, and H. Li, "Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 732–24 741. 18
- [280] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233. 18
- [281] Y. S. Y. Q. M. Zhang, X. L. J. Y. X. Zheng, K. L. X. S. Y. Wu, R. J. C. Fu, and P. Chen, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2021. 18
- [282] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 056–27 087, 2024. 18
- [283] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *arXiv preprint arXiv:2308.02490*, 2023. 18
- [284] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," *arXiv preprint arXiv:2203.10244*, 2022. 18
- [285] C.-Y. Li, K.-J. Chang, C.-F. Yang, H.-Y. Wu, W. Chen, H. Bansal, L. Chen, Y.-P. Yang, Y.-C. Chen, S.-P. Chen *et al.*, "Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation," *Nature Communications*, vol. 16, no. 1, p. 2258, 2025. 18
- [286] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "Plotqa: Reasoning over scientific plots," in *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2020, pp. 1527–1536. 18
- [287] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo, "Chartbench: A benchmark for complex visual reasoning in charts," *arXiv preprint arXiv:2312.15915*, 2023. 18
- [288] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik, "Multimodal learning with graphs," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 340–350, 2023. 18
- [289] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8317–8326. 18
- [290] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209. 18
- [291] J. Wei, N. Xu, J. Zhu, Y. Hao, G. Wu, B. Yu, and L. Wang, "Chartmind: A comprehensive benchmark for complex real-world multimodal chart question answering," *arXiv preprint arXiv:2505.23242*, 2025. 18
- [292] C. Yang, C. Shi, Y. Liu, B. Shui, J. Wang, M. Jing, L. Xu, X. Zhu, S. Li, Y. Zhang *et al.*, "Chartmimic: Evaluating lmm's cross-modal

- reasoning capability via chart-to-code generation," *arXiv preprint arXiv:2406.09961*, 2024. [18](#) [22](#)
- [293] P. Lu, L. Qiu, K.-W. Chang, Y. N. Wu, S.-C. Zhu, T. Rajpurohit, P. Clark, and A. Kalyan, "Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning," *arXiv preprint arXiv:2209.14610*, 2022. [19](#)
- [294] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar, "Infographicvqa," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1697–1706. [19](#)
- [295] J. Wei, C. Jia, Q. Chen, H. He, L. Sun, C. He, L. Wu, B. Yu, and C. Tan, "Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions," *arXiv preprint arXiv:2508.03173*, 2025. [19](#)
- [296] A. Anthropic, "The claude 3 model family: Opus, sonnet, haiku," *Claude-3 Model Card*, vol. 1, no. 1, p. 4, 2024. [19](#)
- [297] OpenGVLab, "Internvl-chat-v1.5: An open-source multimodal large language model," <https://huggingface.co/OpenGVLab/InternVL-Chat-V1-5>, 2024, accessed: 2025-09-24. [19](#) [21](#)
- [298] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llavanext: Improved reasoning, ocr, and world knowledge," 2024. [19](#)
- [299] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024. [19](#)
- [300] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 624–11 641, 2023. [19](#)
- [301] Q. Li, Q. Tao, S. Joty, J. Cai, and J. Luo, "Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 552–567. [19](#)
- [302] A. Foss, C. Evans, S. Mitts, K. Sinha, A. Rizvi, and J. T. Kao, "Causalvqa: A physically grounded causal reasoning benchmark for video models," *arXiv preprint arXiv:2506.09943*, 2025. [19](#) [22](#)
- [303] P. Sarkar and A. Etemad, "Vrbench: Exploring long-form causal reasoning capabilities of large video language models," *arXiv preprint arXiv:2505.08455*, 2025. [19](#)
- [304] X. Li, X. Li, S. Hu, K. Huang, and W. Zhang, "Causalstep: A benchmark for explicit stepwise causal reasoning in videos," *arXiv preprint arXiv:2507.16878*, 2025. [19](#)
- [305] C. Li, E. W. Im, and P. Fazli, "Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 723–13 733. [19](#)
- [306] H. Gao, J. Qu, J. Tang, B. Bi, Y. Liu, H. Chen, L. Liang, L. Su, and Q. Huang, "Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation," *arXiv preprint arXiv:2503.19622*, 2025. [19](#)
- [307] Y. Wang, Y. Wang, D. Zhao, C. Xie, and Z. Zheng, "Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models," *arXiv preprint arXiv:2406.16338*, 2024. [19](#)
- [308] K. Sung-Bin, O. Hyun-Bin, J. Lee, A. Senocak, J. S. Chung, and T.-H. Oh, "Avhbench: A cross-modal hallucination benchmark for audio-visual large language models," *arXiv preprint arXiv:2410.18325*, 2024. [19](#)
- [309] P. Ding, J. Wu, J. Kuang, D. Ma, X. Cao, X. Cai, S. Chen, J. Chen, and S. Huang, "Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 707–10 715. [19](#)
- [310] J. Zhou, Y. Shu, B. Zhao, B. Wu, Z. Liang, S. Xiao, M. Qin, X. Yang, Y. Xiong, B. Zhang *et al.*, "Mlvu: Benchmarking multi-task long video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 691–13 701. [19](#)
- [311] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206. [19](#)
- [312] Y. Weng, M. Han, H. He, X. Chang, and B. Zhuang, "Longvlm: Efficient long video understanding via large language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 453–470. [20](#)
- [313] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 108–24 118. [20](#)
- [314] Y. Shao, Y. Jiang, T. A. Kanell, P. Xu, O. Khattab, and M. S. Lam, "Assisting in writing wikipedia-like articles from scratch with large language models," *arXiv preprint arXiv:2402.14207*, 2024. [20](#)
- [315] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024. [20](#) [21](#)
- [316] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Llavavideo: Video instruction tuning with synthetic data," *Transactions on Machine Learning Research*, 2025. [20](#)
- [317] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023. [20](#)
- [318] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can pre-trained vision and language models answer visual information-seeking questions?" *arXiv preprint arXiv:2302.11713*, 2023. [20](#)
- [319] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, "Socialqa: Commonsense reasoning about social interactions," *arXiv preprint arXiv:1904.09728*, 2019. [20](#)
- [320] S.-Y. Ji, M.-K. Kim, and H.-J. Jun, "Emotion analysis ai model for sensing architecture using eeg," *Applied Sciences*, vol. 15, no. 5, p. 2742, 2025. [20](#)
- [321] Z. Han, B. Zhu, Y. Xu, P. Song, and X. Yang, "Benchmarking and bridging emotion conflicts for multimodal emotion reasoning," *arXiv preprint arXiv:2508.01181*, 2025. [20](#) [22](#)
- [322] T. Zhou, D. Chen, Q. Jiao, B. Ding, Y. Li, and Y. Shen, "Human-vcbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data," *arXiv preprint arXiv:2412.17574*, 2024. [20](#)
- [323] S. Raza, A. Narayanan, V. R. Khazaie, A. Vayani, M. S. Chettiar, A. Singh, M. Shah, and D. Pandya, "Humanibench: A human-centric framework for large multimodal models evaluation," *arXiv preprint arXiv:2505.11454*, 2025. [20](#)
- [324] A. Zadeh, M. Chan, P. P. Liang, E. Tong, and L.-P. Morency, "Social-iq: A question answering benchmark for artificial social intelligence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8807–8817. [20](#)
- [325] M. Kwon, H. Hu, V. Myers, S. Karamcheti, A. Dragan, and D. Sadigh, "Toward grounded commonsense reasoning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5463–5470. [20](#)
- [326] D. Romero, C. Lyu, H. A. Wibowo, T. Lynn, I. Hamed, A. N. Kishore, A. Mandal, A. Dragonetti, A. Abzaliev, A. L. Tonja *et al.*, "Cvqa: Culturally-diverse multilingual visual question answering benchmark," *arXiv preprint arXiv:2406.05967*, 2024. [20](#)
- [327] Z. Ren, J. Ortega, Y. Wang, Z. Chen, Y. Guo, S. X. Yu, and D. Whitney, "Veatic: Video-based emotion and affect tracking in context dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4467–4477. [20](#)
- [328] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng *et al.*, "Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models," *arXiv preprint arXiv:2501.16566*, 2025. [20](#)
- [329] J. Hu, H. Shi, C. Dai, Z. Li, P. Song, and M. Wang, "Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark," *arXiv preprint arXiv:2508.16859*, 2025. [20](#)
- [330] Y. Fu, Q. Liu, Q. Song, P. Zhang, and G. Liao, "Multi-hm: A chinese multimodal dataset and fusion framework for emotion recognition in human–machine dialogue systems," *Applied Sciences*, vol. 15, no. 8, p. 4509, 2025. [20](#)
- [331] Z. Cheng, Z.-Q. Cheng, J.-Y. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110 805–110 853, 2024. [20](#)
- [332] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. [21](#)
- [333] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, "Videollama 2: Advancing

- spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024. 21
- [334] F. Zhang, Z. Cheng, C. Deng, H. Li, Z. Lian, Q. Chen, H. Liu, W. Wang, Y.-F. Zhang, R. Zhang *et al.*, "Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models," *arXiv preprint arXiv:2508.09210*, 2025. 20
- [335] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018. 20
- [336] J. Bi, S. Liang, X. Zhou, P. Liu, J. Guo, Y. Tang, L. Song, C. Huang, G. Sun, J. He *et al.*, "Why reasoning matters? a survey of advancements in multimodal reasoning (v1)," *arXiv preprint arXiv:2504.03151*, 2025. 20
- [337] T. Adewumi, L. Alkhalef, N. Gurung, G. van Boven, and I. Pagliai, "Fairness and bias in multimodal ai: A survey," *arXiv preprint arXiv:2406.19097*, 2024. 21
- [338] J. Lu, L. Song, M. Xu, B. Ahn, Y. Wang, C. Chen, A. Dehghan, and Y. Yang, "Atoken: A unified tokenizer for vision," *arXiv preprint arXiv:2509.14476*, 2025. 21
- [339] H. Lin, T. Wang, Y. Ge, Y. Ge, Z. Lu, Y. Wei, Q. Zhang, Z. Sun, and Y. Shan, "Toklip: Marry visual tokens to clip for multimodal comprehension and generation," *arXiv preprint arXiv:2505.05422*, 2025. 21
- [340] T.-H. Pham and C. Ngo, "Multimodal chain of continuous thought for latent-space reasoning in vision-language models," *arXiv preprint arXiv:2508.12587*, 2025. 21
- [341] S. Liu, H. Ye, L. Xing, and J. Zou, "Reducing hallucinations in vision-language models via latent space steering," *arXiv preprint arXiv:2410.15778*, 2024. 21
- [342] N. Jiang, A. Kachinthaya, S. Petryk, and Y. Gandelsman, "Interpreting and editing vision-language representations to mitigate hallucinations," *arXiv preprint arXiv:2410.02762*, 2024. 21
- [343] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024. 21
- [344] Y. Xu, C. Li, H. Zhou, X. Wan, C. Zhang, A. Korhonen, and I. Vulić, "Visual planning: Let's think only with images," *arXiv preprint arXiv:2505.11409*, 2025. 21
- [345] C. Li, W. Wu, H. Zhang, Y. Xia, S. Mao, L. Dong, I. Vulić, and F. Wei, "Imagine while reasoning in space: Multimodal visualization-of-thought," *arXiv preprint arXiv:2501.07542*, 2025. 21
- [346] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei, "Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 90277–90317, 2024. 21
- [347] J. Wu, J. Guan, K. Feng, Q. Liu, S. Wu, L. Wang, W. Wu, and T. Tan, "Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing," *arXiv preprint arXiv:2506.09965*, 2025. 21
- [348] Y. Zhan, Z. Wu, Y. Zhu, R. Xue, R. Luo, Z. Chen, C. Zhang, Y. Li, Z. He, Z. Yang *et al.*, "Gthinker: Towards general multimodal reasoning via cue-guided rethinking," *arXiv preprint arXiv:2506.01078*, 2025. 22
- [349] J. Zhang, Y. Liu, W. Liu, J. Luan, R. Yan *et al.*, "Weaving context across images: Improving vision-language models through focus-centric visual chains," *arXiv preprint arXiv:2504.20199*, 2025. 22
- [350] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, "Mantis: Interleaved multi-image instruction tuning," *arXiv preprint arXiv:2405.01483*, 2024. 22